Data Science in Investment Management

by

Manish Singh

B.Tech., Indian Institute of Technology Delhi (2018)
S.M., Massachusetts Institute of Technology (2020)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by: Manish Singh
Department of Electrical Engineering and Computer Science
July 16, 2023

Certified by: Andrew W. Lo
Charles E. and Susan T. Harris Professor Professor of Finance
MIT Sloan School of Management.

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science.
Chair, Department Committee on Graduate Students

Data Science in Investment Management

by

Manish Singh

Submitted to the Department of Electrical Engineering and Computer Science
on July 16, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, titled "Data Science in Investment Management," we aim to explore the applications of data science and artificial intelligence across various dimensions of investment management, offering innovative solutions and insights to the industry. This thesis is composed of several parts, each addressing a different aspect of investment management and leveraging data science techniques to deliver valuable insights.

In first part, for industries and crypto-currencies, we develop a dynamic classification system that groups stocks according to quantified similarities from a wide variety of structured and unstructured data features. With the availability of big data, we were able to use artificial intelligence (AI) methods to extract relevant information about companies from various data sources and learn about their similarity in the future, according to market perception. In second part, we study ways of creating capital and portfolio management for fusion energy and biopharmaceutical investments. By leveraging computational techniques like portfolio approach, we provide novel insights into the optimal financing strategies for high-risk, high-reward ventures like fusion research and biopharmaceutical investing. We also quantify the impact of clinical trial results on the stock prices of the companies, that can aid biopharma investors in risk management. Given the increasing interest in ESG investing, we study the excess-returns of the ESG investing. We also develop the measure of the impact on patient lives due to the products of the biopharmaceutical companies that can attract ESG funds for biopharmaceutical companies. Next part of the thesis investigates the real-time psychophysiological analysis of financial risk processing, offering a deeper understanding of human behavior in the context of investment decision-making using a data driven approach. In the next part, we focus on the use of explainable Machine Learning for an important problem of consumer credit risk. In the final part, we conclude with the discussion about the future of Artificial Intelligence and Data Science in Finance.

Thesis Supervisor: Andrew W. Lo
Title: Charles E. and Susan T. Harris Professor, a Professor of Finance

MIT Sloan School of Management.

Acknowledgments

Growing up in different small towns of India, travelling out of India was a dream and studying at M.I.T. was something I never thought about. However, these dream came true due to constant support of my parents papa (J.P. Singh) and mummy (Beena Singh). They motivated, inspired, guided me. Words are not enough to convey my gratitude. Along with my parents, my brother (Robin) and sister (Pallavi) have been a constant source of guidance and I can rely for everything on you both.

I am utterly grateful to Professor Andrew W. Lo for his guidance, opportunities of working on multiple interesting impactful and practical projects over time. Its been a privilege to meet him regularly and an honour to be a teaching assistant for his classes. My learning experience working with Professor Lo just cannot be quantified.

I would like to thank my thesis committee, Professor Munther A. Dahleh and Professor Roberto Rigobon for the valuable guidance, feedback and evaluating my work. It's been a privilege to collaborate with Professor Rigobon on multiple projects throughout the Ph.D.

I am thankful to everyone in Laboratory for Financial Engineering (LFE): especially Jayna Cummings:- for managing all the projects, any administrative task and most importantly correcting all my research articles. Similarly, I am thankful to Kate Lyons, Robert Barron, Mavanee Nealon and Crystal Myler for administrative support, social outings and everything else they have done to support the research. Many thanks to Zied Ben Chaouch, Kien Wei Siah, Chi Heem Wong, Joonhyuk Cho for insightful discussions, assistance with research and joint projects. Apart from being colleagues, you all have been great friends, restaurants partners and tennis/badminton/skii partners.

I am grateful to all the undergraduate research assistants: Abilash Prabhakaran, Roland Rocafort, Aaliya Hussain, Angelina Xu, Cathy Cai, Allison Li, Sarah Wang, Kimberley Dutta, Jessie Y Liu. Working with all of you was very productive for the projects and and up-to-date about the undergraduate life of M.I.T.

I am also thankful to the collaborators and co-authors, who helped at multiple stages in the research: Ruixun Zhang, Florian Berg, Mohammad M. Ghassemi, Peter Zangari, Xinxin Wang, George Bonne, Howard Zhang, Zachery M. Halem, GPT 3.5/4, Arash Nourian, Randall Davis, Sudhanshu Mishra, Nicholas Wu, Abdullah Alhamdan, Irene Hernandez and Dennis Whyte.

I am grateful to my friends in Boston/USA for keeping me sane and for the adventurous travel plans and exploring random activities together.

Contents

1 Introduction **23**

I Peer Grouping **30**

2 Peer Grouping : An Artificial Intelligence-Based Industry Peer Grouping System **31**

2.1 Introduction 32

2.2 Data and Feature Extraction 37

2.3 Distance Metrics 40

2.4 Clustering 41

2.5 Evaluating Performance 42

2.6 Conclusion 60

3 Peer Grouping: Crypto-Connections:- Exploring the Peers **63**

3.1 Introduction 63

3.2 Methodology 65

3.2.1 Data 66

3.2.2 Featurization 68

3.2.3 Hierarchical Clustering 69

3.3 Results 70

3.4 Conclusion 79

II Finance For Impact **80**

4 Finance For Impact: The Reaction of Sponsor Stock Prices to Clinical Trial Outcomes **81**

4.1 Introduction 82

4.2 Literature Review 84

4.3 Data 86

4.4 Methodology 88

4.5 Results and Discussion 91

4.6 Robustness Checks 105

4.7 Conclusion 105

5 Finance For Impact: Megafund **107**

5.1 Drug Development 107

5.1.1 Method 109

5.1.2 Results 111

5.2 Fusion Energy Research 116

5.2.1 Method 117

5.3 Conclusion 121

6 Finance for Impact: Quantifying the Returns of ESG Investing **125**

6.1 Introduction 125

6.2 Methodology 130

6.2.1 Quantifying Excess Returns 130

6.2.2 Portfolio Creation 132

6.2.3 ESG Rating Aggregation 133

6.3 Data and Summary Statistics 136

6.3.1 Data 136

6.3.2 Summary Statistics: Excess Returns and ESG 137

6.4 Performance of ESG Portfolios 139

6.4.1 Correlations between ESG ratings and Excess Returns 139

6.4.2 Returns and Factor Exposure . . . 141
6.4.3 Estimating Excess Returns . . . 143
6.4.4 Treynor-Black Portfolios . . . 147
6.4.5 Combining ESG and Passive Portfolios . . . 147
6.5 ESG as Univariate Impact Factors . . . 151
6.5.1 Environment Portfolios . . . 151
6.5.2 Social Portfolios . . . 151
6.5.3 Governance Portfolios . . . 151
6.6 Conclusion . . . 155

7 Finance for Impact: From Pills to Power: Unveiling the Impact of BioPharmaceutical Companies 157
7.1 Introduction . . . 157
7.2 Methodology . . . 160
7.2.1 Data . . . 160
7.2.2 Pipeline . . . 161
7.3 Analysis of Scores . . . 163
7.4 Conclusion . . . 168

III Behavioural Finance 170

8 Behavioural Finance: Real-time extended psychophysiological analysis of financial risk processing 171
8.1 Introduction . . . 172
8.2 Literature review . . . 175
8.3 Materials and methods . . . 177
8.4 Results . . . 183
8.5 Discussion . . . 188
8.6 Conclusion . . . 191

IV Explainable Artificial Intelligence (xAI) 193

9 xAI : Explainable Machine Learning Models of Consumer Credit Risk 195

9.1 Introduction . . . 195

9.2 Literature Review . . . 200

9.3 Machine Learning Methods . . . 202

9.3.1 Random Forests . . . 203

9.3.2 Inductive Logic Programming . . . 204

9.3.3 Optimal Classification Trees . . . 205

9.3.4 Neural Networks . . . 206

9.4 Data . . . 206

9.5 Evaluation . . . 208

9.6 Explaining Machine Learning Models . . . 211

9.6.1 Interpretability for Loan Companies: Opening the Black Box . 212

9.6.2 Interpretability for Regulators: Model Fairness and Stress Testing 221

9.6.3 Interpretability for Loan Applicants: Counterfactual Suggestions 223

9.6.4 Interpretability for Researchers and Data Scientists: Simple Rules to Summarize the Dataset . . . 229

9.7 Conclusion . . . 231

V Evolution of NLP 233

10 From ELIZA to ChatGPT: The Evolution of NLP and Financial Applications 235

10.1 Introduction . . . 235

10.2 Evolution of NLP Models . . . 238

10.3 Applications . . . 248

10.4 What's Next . . . 256

10.5 Conclusion . . . 258

VI Conclusion 260

A Peer Grouping : An Artificial Intelligence-Based Industry Peer Grouping System 263

B Peer Grouping: Crypto-Connections:- Exploring the Peers 277

C Finance For Impact: The Reaction of Sponsor Stock Prices to Clinical Trial Outcomes 281

C.1 Data 281

C.2 Results: Abnormal Returns Over Time 282

D Appendix Megafund 293

D.1 Parameters of Simulation 293

D.2 Sensitivity Analsysis 296

E Finance for Impact: Quantifying the Returns of ESG Investing 301

E.1 Rank Autocorrelation 301

E.2 Implied Realized Alphas 301

E.3 Sharpe Ratios of Combined Market and Treynor-Black ESG Portfolios 302

F Behavioural Finance: Real-time extended psychophysiological analysis of financial risk processing 305

F.1 Market Efficiency 305

F.2 Physiological Data Collection 306

F.3 Market and Financial Data Collection 306

F.4 Trader Survey 307

F.5 Data Preprocessing 307

F.6 Activation Feature Extraction 310

F.7 Traders Feedback Session 311

F.8 Trader's Activation and Financial Product Type 312

G xAI : Explainable Machine Learning Models of Consumer Credit Risk **321**

G.1 Dataset Details . . . 321

G.1.1 Glossary of Relevant Credit Modeling Terms . . . 321

G.1.2 Explanation of Predictor Features . . . 322

G.1.3 Data Cleaning . . . 324

G.1.4 Data Pre-processing . . . 325

G.1.5 Data Visualization . . . 326

G.2 Model Implementation Details . . . 326

H From ELIZA to ChatGPT: The Evolution of NLP and Financial Applications **331**

H.1 Complete ChatGPT Interaction . . . 331

List of Figures

1-1 Evolution of computation and number of datapoints for AI systems training. Dataset is obtained from [253] 24

2-1 Flow Chart of Methodology 37

2-2 Predictions of eBay Peers for 2014. 55

2-3 Feature Importance. 56

2-4 Feature Importance Aggregated by Dataset. 56

2-5 Examples from the Cluster Visualization Framework. 59

3-1 Number of Coins with Whitepapers in our Dataset. 67

3-2 Number of Coins with Returns in our dataset. 67

3-3 Number of Coins in top x%age of market capitalization. 68

3-4 Distribution of Bitcoin Similarity with other Currencies 71

3-5 Time Varying β for the Whitepaper Similarity Score. 72

3-6 Time Varying β for the Return Similarity Score. 74

3-7 Time Varying β for the Fundamental Similarity Score. 75

3-8 Clustering of Cryptocurrencies based on the Return Similarity for Year 2021. 78

3-9 Log Cumulative Returns of the Bitcoin Cluster Factor 79

4-1 Steps involved in this event study, from the initial data collection to the analysis of our results. 85

4-2 Timeline of the event. 90

4-3 Left: Safety and Adverse Effect , Right: Lack of Efficacy 94

4-4 Left: Early Positive Outcome, Right: Primary Endpoints Met 94
4-5 Primary Endpoints not met 94
4-6 Relative abnormal returns due to different clinical trial and sponsor company factors on the day of event (day 0-1). 104

5-1 Flowchart depicting the megafund simulation framework. 109
5-2 Number of projects and investment across phase and therapeutic area for benchmark and optimized portfolio. 112
5-3 Flowchart depicting the megafund simulation framework. 118

6-1 The correlation between average ESG score & stock performance (alpha) (y-axis) versus time (x-axis). 140
6-2 Realized alphas for portfolios created using ESG scores (y-axis) versus implied alphas (x-axis) for the same portfolios using the methods described in Section 6.2.1. 145
6-3 Annualized returns (y-axis) versus volatility (x-axis). 150

7-1 Bass Model. x-axis: year from launch, y-axis: f(t), number of new adopters. 163
7-2 Histograms of Impact Scores of the Companies. Left: Raw Score, Right: Log Transformed Scores. x-axis: Impact Scores, y-axis: Number of Companies. 165
7-3 y-axis: Percentage of Impact Contributed by the Disease Category. x-axis: Disease. 166

8-1 Histograms of four PP activation metrics. 184
8-2 Evolution of trader's PP activation around a transaction. 189

9-1 LIME Approximation 216
9-2 Feature Importance LIME 217
9-3 Feature Importance SHAP 219
9-4 Extreme Scenarios: LIME 224
9-5 Optimal Trees Model 230

10-1 Word cloud representing the most frequently occurring words in this article. . . . 237
10-2 Evolution of computation and number of datapoints for AI systems training. Dataset is obtained from [253] . . . 239
10-3 Parsing Tree. . . . 241
10-4 The word embedding vectors (obtained using embedding-based NLP method word2vec) plotted on 2-D plane. . . . 244
10-5 Summary of the different steps in the pipeline of a large language model. 245
10-6 Topic Modelling Risk Factors. . . . 250
10-7 Three examples of NLP application to climate-related impact investing. 252
10-8 Out-of-sample cumulative log returns of portfolio sorted based on sentiment score constructed from news articles. . . . 254
10-9 Event Study: Stocktwit and Twitter. . . . 255

A-1 Yearwise Average Pairwise Correlation . . . 273
A-2 Standard Deviation fo Cluster Sizes. . . . 275
A-3 Feature Importance Datasets. . . . 275

C-1 Data Cleaning Process. . . . 282
C-2 Literature Review Summary. . . . 283

D-1 Block correlation matrix . . . 296
D-2 Discounted Cash Flow analysis of royalty investement . . . 296

F-1 E4 Wristband. . . . 315
F-2 Inactive Periods Detection. . . . 316
F-3 Examples of rolling correlation between two physiology signals averaged over all traders. (A) EDA mean vs. EDA slope; (B) EDA mean vs.HR RMSSD; (C) EDA mean vs. HR mRRi; (D) EDA mean vs. BVP mean. . . . 316
F-4 Granger Causality Test Histograms . . . 319

G-1 Household Debt. . . . 322

G-2 23 x 23 Correlation Matrix. 327

List of Tables

2.1 Datasets and Various Features Extracted 40

2.2 Evaluation of ML models. 44

2.3 Results after hierarchical clustering. 46

2.4 Adjusted Rand Index for ML Clusters. 47

2.5 Cluster Stability. 47

2.6 Diversification Table for Cluster Factor Returns. 49

2.7 Intra-Cluster Dispersion for Various Fundamental Factors. 51

2.8 Diversification Table. 52

2.9 Hedging Risk Exposures. 53

3.1 The list of fundamental factors computed for each crypto-currency coin. 69

3.2 Volatility (monthly in percentage) of Coins and Long-Short Portfolio Constructed Using Peers 76

3.3 Mean Returns and Sharpe Ratio of Equal Weighted Cluster Factor Portfolios and Market Index. 77

4.1 Average abnormal returns for different outcome types on various significant dates 92

4.2 Outcome types 102

4.3 Target Accrual 102

4.4 Company sponsor category 102

4.5 Phase categories. 102

4.6 Disease therapeutic category 102

4.7 Clinical Trial Design 103

4.8 Regression Constant 103

5.1 114

5.2 Performance statistics of the megafund. 119

5.3 Average number of spin-off companies from the successful megafund. . 120

5.4 Megafund Performance. 121

5.5 Megafund Sensitivity Analysis. 122

6.1 The cross-sectional rank correlation between ESG scores (individual and aggregate) averaged over time. 138

6.2 Alpha Statistics. 138

6.3 Average cross-sectional correlation between ESG scores and alpha (excess returns). 139

6.4 ESG Portfolio Performance 142

6.5 Factor Exposure 143

6.6 Excess Returns for ESG Portfolios 146

6.7 Excess Returns for Treynor Black ESG Portfolios 148

6.8 Excess Returns for E Portfolios 152

6.9 Excess Returns for S Portfolios. 153

6.10 Excess Returns for G Portfolios. 154

7.1 Informa Data Summary for Launched Drugs. 160

7.2 GBD Data Summary. 161

7.3 Companies with High Impact Scores 164

7.4 Companies with High Market Capitalization Normalized Impact Scores 165

7.5 Drugs with High Impact Scores 167

7.6 Correlation between Scores. *: p-val < 0.05 and **: p-val < 0.01 . . 168

7.7 Correlation between Scores and Returns. *: p-val < 0.05 and **: p-val < 0.01 168

8.1 Summary statistics of four PP activation metrics. 183

8.2 Granger Causality Test 187

9.1 5-Fold cross-validation performance of ML models. 210
9.2 Constraints Violation 215
9.3 Extreme Customers (data points). 222
9.4 Counterfactual 227
9.5 Counterfactual Creditworthy 228
9.6 Proximity Analysis. 228

10.1 Lexicon and topic-modeling examples of NLP 242

A.1 Datasets used for feature extraction and the number of companies covered in each dataset every year. 264
A.2 Description of different factors and their speed of change. 266
A.3 Train-validation-test year split. 269
A.4 Evaluation of Cosine Model. 270
A.5 Results after hierarchical clustering: Cosine Model. 270
A.6 Pairwise Return Correlation: Yearwise. 272
A.7 Return Correlation Standard Errors. 274

B.1 ID Name Mapping of the Cryptocurrencies. 280

C.1 Number of Trials 284
C.2 Regression Coefficient: Time Series. 285
C.3 Outcome types 286
C.4 Target Accrual 286
C.5 Company sponsor category 286
C.6 Phase categories. 286
C.7 Disease therapeutic category 286
C.8 Clinical Trial Design 287
C.9 Regression Constant 287
C.10 Outcome types 288
C.11 Target Accrual 288
C.12 Company sponsor category 288

C.13 Phase categories. 288
C.14 Disease therapeutic category 288
C.15 Clinical Trial Design 289
C.16 Regression Constant 289
C.17 Outcome types 289
C.18 Target Accrual 290
C.19 Company sponsor category 290
C.20 Phase categories. 290
C.21 Disease therapeutic category 290
C.22 Clinical Trial Design 291
C.23 Regression Constant 291

D.1 Probability of successful phase transition 293
D.2 Cost of a clinical trial 294
D.3 Duration of clinical trials 294
D.4 Market value of drugs 295
D.5 297
D.6 Portfolio performance varying Probability of Success 298
D.7 Market Value of Drugs 299
D.8 Portfolio performance on varying market value of drugs. 300

E.1 Cross-sectional rank autocorrelation between vendor ESG scores and aggregate scores. 302
E.2 Model-implied and Realized Alphas 303
E.3 performance of Market+ Treynor Black ESG Portfolios. 304

F.1 Physiological waveforms measured by the Empatica E4 wristband for each trader. 313
F.2 Market time series data acquired from different data sources. 313
F.3 Total number of transactions made by the traders in each week. . . . 313
F.4 Aggregating Transactions. 314

F.5 Correlation Between Features. 317

F.6 Correlation Between PP Signals. 318

F.7 Regression coefficient of trader activation in the column business division with row business division as the baseline and the associated p-value. 318

G.1 Five-fold cross-validation data distribution. $y_i = 1$ corresponds to the data point (i.e., individual) who defaulted. The dataset is fairly balanced. 328

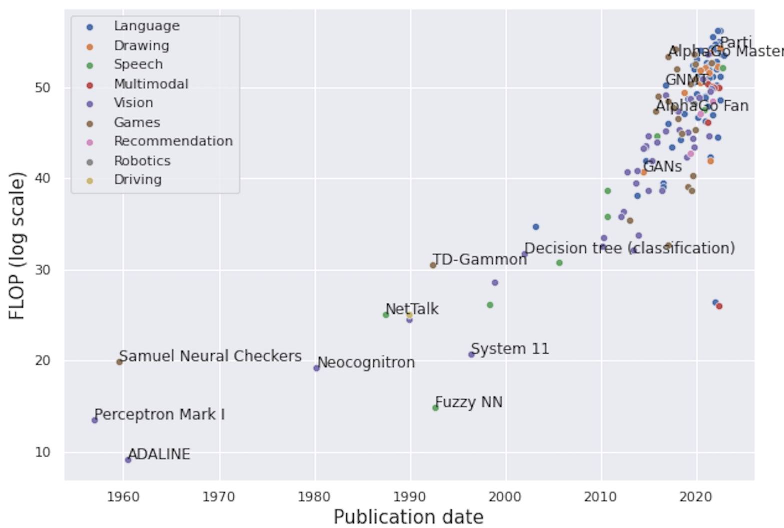
Chapter 1

Introduction

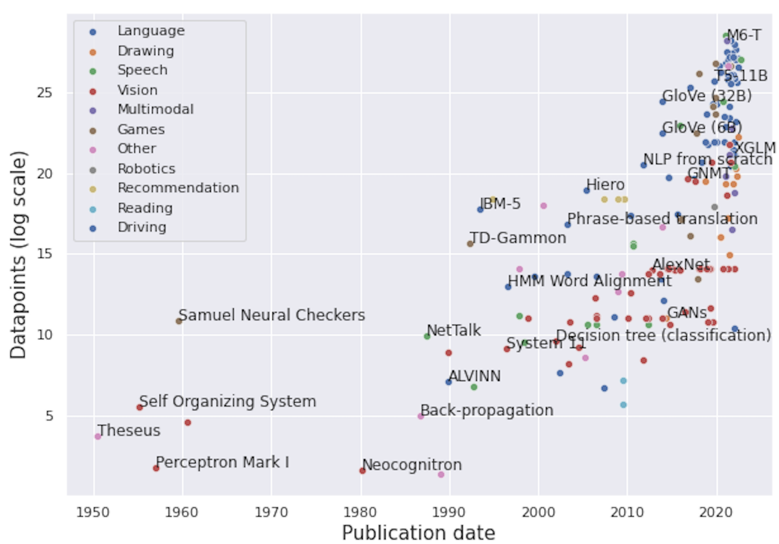
Artificial Intelligence (AI) models have evolved over time with theoretical innovation and the availability of data and computational resources. Figure 1-1 shows the computation and datapoints used by the artificial intelligence (AI)–including language systems– systems developed over time where we can observe an exponential increase in the computation and datasets used by AI systems. This was possible due to the increase in the computational resources and decreasing cost of computational and data storage facilities.

The rapid advancements in data science and artificial intelligence have significantly transformed the landscape of investment management in recent years. The availability of data and AI systems has revolutionized the world of investment management and finance in several significant ways. These technologies have enabled investors, fund managers, and financial institutions to make more informed decisions, automate processes, and uncover hidden patterns in the market. This thesis, titled "Data Science in Investment Management," aims to explore the potential applications of data science and artificial intelligence across various dimensions of investment management, offering innovative solutions and insights to the industry.

The thesis is composed of several chapters, each addressing a different aspect of investment management and leveraging data science techniques to deliver valuable insights. The first two chapters (Part I) focus on artificial intelligence-based peer grouping systems for both traditional industries and cryptocurrencies. It has broad



(a) Computation used to train notable artificial intelligence systems.



(b) Number of datapoints used to train notable artificial intelligence systems.

Figure 1-1: Evolution of computation and number of datapoints for AI systems training. Dataset is obtained from [253] application because investors and financial analysts use it for investment management and competitive industrial diligence. Currently, multiple industry peer grouping system exists including North American Industry Classification System, the Global Industrial Classification System, and the Industry Classification Benchmark. However, they are static (doesn't evolve over time) and doesn't incorporate market perception (defined as return correlation).

For industries, we develop a dynamic industry classification system that groups stocks according to quantified similarities from a wide variety of structured and unstructured data features. With the availability of big data, we were able to use artificial intelligence (AI) methods to extract relevant information about companies from various data sources and learn about their similarity in the future, according to market perception. Our AI-driven Peer Grouping System (AIPGS) adapts itself and evolves over time in tandem with the changing perceptions of companies.

Contrary to the companies in the equity market, the crypto-currency market is new and currently the existing crypto currency groupings (taxonomy) are not tailored to the application of investment management. Hence, we create peer grouping system for crypto-currencies using various data sources including: historical returns, white papers, fundamental factors and fundamental properties.

The part II of thesis delve into the energy, pharmaceutical sectors and ESG investing. Investment in energy, healthcare and pharmaceutical sectors is crucial due to several reasons that impact society, the economy, and individual well-being. These sectors play a vital role in improving and maintaining public health, driving innovation, and contributing to economic growth. However, there is dearth of funding in fusion energy and pharmaceutical sector. Hence, in different chapters we study ways of creating capital and portfolio management for fusion energy and biopharma investments.

For biopharmaceutical companies, apart from general factors like earnings announcements, dividends, corporate actions; clinical trial outcomes are critical since they play a major role in determining the future revenue, growth, and valuation of the company sponsoring the trial. For investors in this sector, clinical trial results are the predominant factor determining their return on investment. Therefore, quantifying the impact of clinical trial results on the stock prices of the companies is of great financial significance. In one of the chapter, we investigate for the factors responsible for varying abnormal returns by performing an event study analysis for 13,807 clinical trial outcomes using data from 2000 to 2020 for 379 publicly traded U.S. companies.

We benefit from the availability of the big data (in context of bio-pharmaceutical domain) to accurately investigate for the factors and develop model that can quantify the relationship between the clinical trial factors and abnormal returns of companies. The model can be used by investment managers to obtain average abnormal returns for the stock of a sponsor company at the declaration dates of a clinical trial result.

The following chapters address various aspects of investment project financing, diversified drug development projects along with the fusion energy. By leveraging computational techniques like portfolio approach, these chapters provide novel insights into the optimal financing strategies for these high-risk, high-reward ventures.

Final chapters of part II focuses on sustainable investing. Over recent years, impact investing have gained a lot of attention from investors. There have been an estimate of $9 trillion impact investing market in the US alone. According to united nations principles of responsible investing, number of organisations that are signatories of UNPRI have increased from 450 in 2016 to 4935 in 2022 representing 1.3$ trillion asset under management in 2016 to over 100$ trillion in 2022.

Given the increasing interest in ESG investing, investors are interested in knowing excess-returns of the ESG investing (if any). Hence, we develop the portfolios using ESG scores from different rating agencies to quantify the excess returns due to ESG. Next, due to varying methodologies of vendors for the computation of the ESG scores, the divergence of the ratings are observed across vendors. This divergence implies, the ESG scores can be considered to be composed of ESG signal and noise. Due to low signal-to-noise ratio, it is hard to estimate the true ESG excess-returns.

To solve for problem of noise, we propose different ways of aggregation of the ESG scores across vendors. We construct a single measure of ESG by combining ESG scores from six different vendors using various statistical and voting aggregation techniques.

Since the aggregate scores are less noisy, the estimates of various properties are more representative of the true properties of ESG portfolios. It can give an historical perspective of ESG performance for the investors. Hence, we use data science (data from rating agencies) to solve for the problem of noise.

In the final chapter of part II, we develop the measure of the impact on patient lives due to the products of the biopharmaceutical companies. The current ESG score methodologies doesn't include the contribution of the drugs/treatments developed by the biopharma companies in the computation of the ESG scores. Biopharma companies are creating "social impact" by developing treatments and saving the lives of the patients. The social impact measure can be used by investors for the impact investing in biopharmaceutical sector. Which in-turn can generate capital for drug development.

Part III investigates the real-time psychophysiological analysis of financial risk processing, offering a deeper understanding of human behavior in the context of investment decision-making.

Traders are highly trained and motivated to be "rational". Given that trades may not be successful all the times, the fluctuating profit and loss (PnL) may lead to incidences of stress. Hence, in financial markets, sudden fluctuations in prices, changes in investor holdings, or other factors unknown to the traders may impact decisions. There have been incidences of rogue traders, for example, Nick Leeson 1 Billion$ loss led to fall of Barings bank in 1995, Jerome Kerviel's illegal trades caused 4.9 Billion Euros loss to Societe Generale.

The primary focus of our analysis is to measure the relation (if any) between a trader's PP state and their financial decisions or market events, both contemporaneously and temporally. By analyzing the changes in the trader's PP activation levels, we identify the circumstances in which such changes coincided with, or were caused by, financial transactions or market events. This information can be used to give feedback to the traders and prevent them from making irrational decisions. This is one of the largest study, where we collected the physiological signals of 55 professional financial traders at a global financial institution during the entire trading day over a five-day period for each trader. The data-driven approach allowed us to study and obtain significant relationships.

In Part IV, we focus on the use of Machine Learning for an important problem of consumer credit risk. The total U.S. household debt at the end of the fourth quarter of 2020 is estimated to be $14.56 trillion, with 189.6 million new credit accounts opened within the prior 12 months. Given this massive amount of household debt, even a low delinquency rate can significantly affect the operation of the financial system. This potential impact makes the study of credit default risk an important real-world classification task.

In this study, we identify the different stakeholders involved in credit risk management—loan companies, regulators, loan applicants, and data scientists—and provide explanations of the models according to their needs. We modified the xAI models for their use in credit risk management. Our study demonstrates that the functionality of black-box ML models can be explained to a range of different stakeholders, if the right tools are applied to the task. These progresses unlock the future potential of applying AI to improve credit modeling.

Finally in part V, we study the evolution of NLP over decades: from ELIZA (the first chatbot) to GPT 3.5/4 (the state-of-art). The use of NLP in finance has evolved significantly over the past few decades. Initially, NLP was used in finance for comparatively simple tasks, such as sentiment analysis of financial news articles and stock predictions. However, with advances in NLP techniques and increased computational power, the application of NLP to finance has become more sophisticated and diverse. Today, NLP models are used in finance for a wide range of tasks, including asset management, risk management, and impact investing.

However, it is important to consider the limitations of NLP, such as the problem of data scarcity, its ethical and legal implications, and the inherent limitations of NLP models themselves. Another important limitation to consider is the apparent "hallucinations" of AI bots which can lead to the strange and sometimes unsettling experiences for users.[1] Overall, the future of NLP in finance is bright, but it will likely

[1] https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

be accompanied by a combination of technological pitfalls and practical challengesas the financial industry adapts to these newfound powerful capabilities.

In summary, the availability of data and AI systems has transformed the world of investment management and finance by enabling better decision-making, increased efficiency, and the discovery of new opportunities. It has allowed to generate capital for the fusion energy and pharmaceutical sectors and have solved problems related to ESG and sustainable investing. As these technologies continue to advance, their impact on the industry will only grow, driving further innovation and change.

Part I

Peer Grouping

Chapter 2

Peer Grouping : An Artificial Intelligence-Based Industry Peer Grouping System

In this Chapter, we develop a data-driven peer grouping system using artificial intelligence (AI) tools to capture market perception and, in turn, group companies into clusters at various levels of granularity. In addition, we develop a continuous measure of similarity between companies, and use this measure to group companies into clusters and construct hedged portfolios. In our peer groupings, companies grouped in the same clusters had strong homogeneous risk and return profiles, while different clusters of companies had diverse, varying risk exposures. We extensively evaluated the clusters robustly and found that companies grouped together by our method had higher out-of-sample return correlation but lower stability and interpretability than companies grouped by a standard industry classification system. We also develop an interactive visualization system for identifying AI-based clusters and similar companies.

2.1 Introduction

Industry classification refers to the organization of companies into groups. It has broad application because national agencies need it to summarize economic statistics accurately, and investors and financial analysts use it for investment management and competitive industrial diligence.

Industry classification started in 1937, via the Standard Industrial Classification system (SIC). This system classified companies based on their activities at the establishment level. Since then, multiple industry classification systems have been established, grouping companies by different criteria. For example, the North American Industry Classification System (NAICS) groups companies based on production at the establishment level, the Global Industrial Classification System (GICS®[1]) and the Industry Classification Benchmark (ICB) classify companies according to revenue, earnings and market perception [209], and the Text-Based Network Industry Classifications (TNIC) groups companies according to business descriptions in 10-K reports.

Currently, none of the existing classification systems are solely based on market perception. However, due to the widespread adoption of industry groupings by multiple stakeholders in investment management, including traders, hedge funds, asset management firms, and stock brokers, the public market's perception has become vital in classification systems because it allows stakeholders to efficiently compare a company's financial performance with its peers, and it allows the hedging and construction of diversified portfolios.

In this chapter, we create a dynamic industry classification system that groups stocks according to quantified similarities from a wide variety of structured and unstructured data features. We trained machine learning (ML) models of these features to predict future return correlations between stocks, which we view as a proxy for market perception. We define market perception as investors' views about a company that drive the stock returns' movement and their co-movement with similar stocks.

[1] Global Industry Classification System is jointly owned by MSCI Inc. and S&P Global.

However, future return co-movement is dynamic and difficult to measure. Nevertheless, with the availability of big data, we were able to use artificial intelligence (AI) methods to extract relevant information about companies from various data sources and learn about their similarity in the future, according to market perception. Our AI-driven Peer Grouping System (AIPGS) adapts itself and evolves over time in tandem with the changing perceptions of companies.

Return correlation between companies has been widely used to determine the homogeneity of industry groups as perceived by investors [51, 82]. We observed that clusters using a distance metric calculated via return correlation will group companies with similar return characteristics. We calculated the industry effects, and demonstrated that our ML models group firms together with similar fundamental factor exposures. The clusters obtained by this method enhance the potential benefits that can be gained by diversification based on clusters in the portfolio creation process.

In our dynamic peer grouping system, we applied AI methods at different steps to enhance the system's ability to group together similar firms. Specifically, we first obtained several datasets on returns, factor exposure, news co-mentions, and 10-K filings. From these datasets, we extracted features via various methodologies, including (but not limited to) term frequency-inverse document frequency (TF-IDF) and Doc2Vec for 10-K filings and Node2Vec for news mentions, to capture the potential similarity between firms[2]. We then used ML models such as ridge regression, neural networks, and XGBoost to learn and predict the relationship between features and future return correlation between companies. Finally, we used hierarchical clustering to group similar firms together in sector, industry, and sub-industry groups.

In AIPGS, we learned relationships and found groups that are not captured in existing classification systems. For example, in 2018, one sub-industry group from AIPGS included only Alphabet, Netflix, Apple, Amazon, and Facebook. All of these companies are prominent U.S. technology companies and popularly referred to as *FAANG*. However, GICS would have traditionally classified these companies in dif-

[2]We describe TF-IDF, Doc2Vec, and Node2vec in more details in Section **Data and Feature Extraction**. These are the AI based feature extraction methods from the text and graph datasets.

ferent sub-industries: respectively, as interactive media and services; movies and entertainment; technology, hardware, storage, and peripherals; internet and direct marketing retail; and interactive media and services. Likewise, GICS generally grouped PepsiCo with the Coca-Cola Company and classified PepsiCo as a soft drinks or beverage company at subindustry and industry levels, but AIPGS generally separated PepsiCo from Coca-Cola and grouped PepsiCo in the same subindustry and industry as packaged food or food companies such as General Mills, Mondelez International, and The Hershey Company. This is insightful because PepsiCo earns more revenue from its food/snacks business than from beverages[3].

Similarly, GICS groups Disney in the group movies and entertainment, whereas AIPGS grouped the Walt Disney Company with cable and broadcasting companies such as Comcast Corporation. Disney launched their streaming service Disney+ in 2019, which might potentially explain the grouping. Likewise, Cognizant is an information technology (IT) company that performs data analytics and processing. Although GICS placed Cognizant with IT consulting firms, AIPGS grouped Cognizant with multiple data processing and IT consultancies.

AIPGS also captured groupings that may seem unintuitive or surprising at first glance but can be explained via further fundamental analysis. For example, although eBay may be seen as one of the largest e-commerce sites, AIPGS indicates otherwise. For 2014, the ML model predicted that eBay should have been grouped in the same sub-industry as finance companies such as The Green Dot Corporation, Visa, and Mastercard. Upon further analysis, this classification makes sense because in 2013 eBay owned PayPal, which did not split off until 2015. PayPal's high growth prospects and leadership in the money-transfer space could have caused the market to view eBay as a fintech firm. Interestingly, for 2018, well after eBay had split from PayPal, AIPGS classified eBay in the same sub-industry as e-commerce sites such as Priceline.com, Expedia Group, Etsy, and Shutterstock. This indicates the market's stabilizing perception of eBay as an e-commerce site after its split from PayPal in 2015. Thus, according to our model, for 2014, an investor looking for exposure to

[3] We verified this information via 10-K filings of PepsiCo.

companies with similar stock movements as eBay would not have invested in other popular e-commerce sites, but in fintech and payment processing firms such as Visa and Mastercard.

The popular classification systems, such as NAICS, GICS, and ICB, discretely group companies without providing a continuous degree of similarity within and across groupings. In this work, apart from merely grouping companies, our model produces a continuous measure of similarity, which allows us to find the top k peers of a particular company. We demonstrate that single-stock portfolios can be hedged using peers determined via similarity scores.

This chapter contributes to the literature in data-driven classification systems. [143] proposed a text-based classification system based on business descriptions from 10-K filings and showed that their method outperformed SIC and NAICS in grouping together firms with similar profitability, sales growth, and market risk. Lee (2015) [166] used internet co-searches from the U.S. Securities and Exchange Commission (SEC) website to find groups of similar firms, assuming that users of the SEC data-retrieval website search for similar firms together to aid their investment process. Although this method incorporated investor perception, it is limited by availability of search data. [76] proposed designing an industry classification system from inter-firm transaction data, but failed to produce any industry groupings due to lack of data. [119] and [264] proposed identifying peers from patent citation and stock tweets on Twitter, respectively. Both methods were also limited by availability of data: If a company does not write patents in common with patents of other companies or if a company is not mentioned in stock tweets, then the classification system will fail to assign a grouping. Similarly, [295] proposed a system to quantify company similarity using different data sources (10-K filings, news co-mentions, returns, fundamentals and analyst coverage overlap).

Our analysis builds on the aforementioned methods, and we propose an industry classification model that is explicitly trained to predict the market perception (return correlation) of companies one year into the future. We also incorporated AI methods that led to performance improvements over the baseline. Some of the existing methods suffer from a lack of data availability, and fail to place all public companies into industry groups. Our system is robust because all the datasets used are available (except for news co-occurrence)[4] for almost all publicly traded U.S. companies.[5]

One of our goals is to incorporate a baseline classification system that is robust and popular. GICS is a classification system that is "market-oriented" and used by practitioners. ICB is another market-oriented classification system similar to GICS, as documented in the literature[280]. Other classification systems are not as relevant to investors; e.g., SIC and NAICS are product-oriented classification systems used by federal statistical agencies to analyze data about the U.S. economy. [104] developed an industry classification system that is based on SIC. [32] and [51] compared GICS to NAICS and Fama–French classifications, respectively, and found that GICS more effectively captures stock co-movement, cross-sectional variations in valuation multiples, and other key financial ratios. Hence, we used GICS at various levels of granularity, particularly at the sector, industry, and sub-industry levels, as a benchmark against our classification system. We show that our classification system outperformed GICS: Our system more effectively grouped companies with strong return co-movement and factor exposures.

Next, we show how we calculated a similarity score to group together the most similar companies. These steps are presented in Figure 2-1. First, we calculated the features for all companies from various data sources, as we will describe in Section **Data and Feature Extraction**. We then calculated the similarity score between each pair of companies. The similarity score between the companies was obtained using ML models, as we will describe in Section **Distance Metrics**. With these similarity scores, we determined the distance between every pair of companies and applied hierarchical clustering, as explained in Section **Clustering**, to group all the companies into clusters. Finally, the clusters obtained were evaluated in an out-of-sample period and used for different applications, as we will describe in Section

[4]Our industry classification system is still able to assign an industry grouping if the company is missing in the news dataset.

[5]The companies used in our analysis covered approximately 99% of the free-float-adjusted market capitalization in the U.S.

Evaluating Performance.

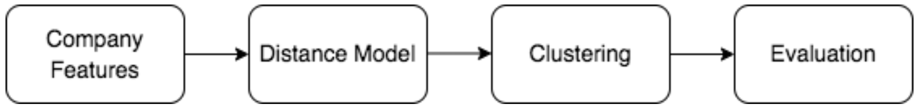


Figure 2-1: Flow Chart of Methodology

2.2 Data and Feature Extraction

We aim to create a classification system that may be useful for investors by incorporating the market's perception of companies. Consequently, we use multiple data sources to capture the similarity between companies. Inspired by MSCI Peer Similarity Scores [295, 217], we use stock returns, factor exposures, 10-K filings, news co-mentions, and the GICS grouping data in our analysis. We next summarize these datasets and feature extraction methods.

Returns

We obtained the daily and monthly returns of all companies. The number of companies and the number of features are described in Table A.1 and in Table 2.1. Our model incorporated returns because companies belonging to the same industry groups experience similar shocks, which are reflected in stock returns. For example, a regulation change in favor of electric vehicles (EV) is favorable to EV companies, which may cause their stock prices to increase, while potentially hurting automobile firms reliant on gasoline or diesel. Also, we used return correlation to quantify the similarity between companies.

Factor Exposures

Multiple factors have been discovered to explain the source of abnormal returns of companies [136]. According to the fundamental factor model [105], companies with a similar factor exposure tend to have similar return characteristics. The factor exposures used in our models are the descriptor exposures from the MSCI Barra US Total Market Model [22] and are listed in the Appendix in Table A.1. It is important to note that, in our model, factor exposures that change slowly over time, such as variability in sales, are driven by inputs that change at an annual or quarterly frequency (slow factors), while others, such as short-term reversal, were designed to have higher turnover (fast factors).

10-K Filings

10-K filings are annual regulatory reports required by the SEC. We use the business description section of 10-K filings, which includes detailed information about companies' businesses and operations. Thus, the filings can be used to generate important features to measure the similarity between companies.

These reports are available in text form. To obtain the features used for the similarity model, we use the following methods:

- **TF-IDF**: TF-IDF is used to determine the most important word in documents. Each document is represented by the vector product of term frequency (the proportion of text the term composes in a document) and the inverse document frequency (the number of documents in which the word occurs) for every word in the document. This method of feature extraction has been used by [189]. Once we calculate the TF-IDF features, we obtain a high-dimensional sparse representation of each document. We reduced the number of dimensions to 100 using truncated singular value decomposition (SVD).

- **Doc2Vec**: This approach involves obtaining a document embedding, as explained in [75]. Doc2Vec is an unsupervised learning method, and using the Distributed Bag of Words and Distributed Memory models, we obtained two representations of every 10-K document, Doc2Vec0 and Doc2Vec1. For the Doc2Vec method, we used a publicly available implementation [6] and made appropriate modifications.

[6] https://markroxor.github.io/gensim/static/notebooks/doc2vec-wikipedia.html

News Co-Mentions

From Ravenpack, we obtained information about companies mentioned together in news articles. Companies appearing together in news articles can indicate similarity. For example, Facebook and Twitter can be considered similar because they are often mentioned together in articles about social media.

Using the news dataset, we extracted two types of features for companies:

- **Co-occurrence**: For every pair of companies, the Jaccard similarity is calculated, as follows: Let A be the news articles with mention of company i and B be the set of news articles with mention of company j. We use $A \cap B / A \cup B$ as their measure of similarity.

- **Node2Vec**: The co-occurrence method fails to capture higher-level connections between companies. For example, if companies i and j are frequently mentioned together in news articles, and if j and k are mentioned together in others, then companies i and k should be similar to some extent even if i and k are never or infrequently mentioned together. Hence, we extract a set of features based on Node2Vec[133]. Node2Vec is a graph-embedding model used to determine the transitive relationships between objects. Therefore, we constructed a graph in which companies are nodes and the co-mentioned companies are connected with the edges. The edges were also weighted by the number of times a pair of companies were mentioned together in an article. Using Node2Vec, we obtained a 128-dimensional embedding of the companies.

GICS

As mentioned earlier, GICS is an industry-standard hierarchical classification system used by financial firms and investors globally. Within each hierarchy, every public company is classified into a sector, industry group, industry and sub-industry. It is a market-oriented classification system that uses revenue as a key factor in determining a firm's principal business activity, along with earnings and market perception. Apart from using GICS as a baseline classification system, we also used it as one of the inputs for the similarity prediction model.

Dataset-Features	Number of Features
Returns-Daily	252
Returns-Monthly	12
Factors	264
10K-TF-IDF	82606
10K-TF-IDF(SVD)	100
10K-Doc2Vec0	200
10K-Doc2Vec1	200
News-Node2Vec	128
GICS	4

Table 2.1: Datasets and Various Features Extracted

2.3 Distance Metrics

Given these features of companies, we used cosine and ML metrics to calculate the similarity scores between companies. These models processed the data available at the end of year t (the in-sample period) to predict and estimate the distance (based on return correlation) between the companies in year $t+1$ (the out-of-sample period). This helps as we seek to avoid any look-ahead bias in the similarity score and distance calculation. Details about the cosine distance model of similarity prediction are available in the Appendix in the *Cosine Distance* section.

We used ML models to predict the out-of-sample distance and similarity between the companies. The ML model took the similarity feature vector between a pair of companies as an input. The similarity feature vector was constructed by appending the L1 distance between different feature vectors with the cosine similarity. It is constructed as follows: Given f_i^k and f_j^k, we constructed $F_{i,j} = \bigcup_k \{|f_i^k - f_j^k|\ ,\ cos(f_i^k, f_j^k)\}$. Here, f is a feature vector, and i and j are companies. However, for our ML models, k may represent different datasets: monthly returns, daily returns, 10-K-TF-IDF(SVD), 10K-Doc2Vec, or news Node2Vec features. We also appended the news co-occurrence to $F_{i,j}$. Finally, the ML model took $F_{i,j}$ as an input and predicted the future similarity between the companies. We calculated future similarity between the companies using future daily returns correlations.

The different ML models we used for this analysis were ridge regression, neural networks, and XGBoost. Details of the ML models are included in Appendix in the *Machine Learning Models* section. The training data consisted of $F_{i,j}$ between all pairs of companies in the sample. For testing year $t+1$ (out of sample), training was performed on year $t-1$, and validation was performed on year t. Once the ML model was trained and hyperparameters were tuned, we used it to obtain the similarity scores between all the pairs of companies for the out-of-sample year $t+1$. After that we performed an evaluation using the predicted similarity score for out-of-sample period $t+1$. We provide further details about the dataset split in the Appendix in Table A.3.

2.4 Clustering

Once we have the similarity scores between the companies for each year, we can cluster them into groupings of similar companies. These groupings should have 1) homogeneity, in which similar companies should be assigned to the same clusters, and 2) fair balance, in which a majority of the companies should not be assigned to a minority of clusters.

Existing industry classification systems, such as NAICS, GICS and ICB, are hierarchical. The hierarchical structure of the financial markets is also validated by [204] and [76]. Hence, we apply hierarchical clustering via an agglomerative (bottom-up) approach. This approach considers all companies as individual clusters. At every subsequent step, the clusters are merged to move up the hierarchy. Hierarchical clustering requires 1) the distance between the pair of observations, and 2) the linkage, which is the method used to merge the clusters upwards in the hierarchy. The similarity score between the pairs of observations is available as described in Section **Distance Metrics**. We calculated the distance as $(1 - SimilarityScore)$.[7] For the linkage, we

[7] $sqrt(2 * (1 - SimilarityScore))$ is used in literature to convert correlation to distance. We

applied Ward's method, which minimizes the intra-cluster variance (i.e., the sum of squares from the center) when two clusters are merged. If a large number of companies are assigned to the same cluster, then the intra-cluster variance increases. Hence, using Ward's method helped the models develop similar-sized clusters.

Using hierarchical clustering, we created groupings that corresponded to sectors, industries, and sub-industries. For a fair comparison with GICS, we approximately matched the number of clusters at each level to GICS groupings, such that if GICS had 11 sectors, then the number of clusters at the sector level for hierarchical clustering was set to 11. However, hierarchical clustering can be used to obtain an arbitrary number of clusters.

2.5 Evaluating Performance

We aim to group stocks based on the perception of the market and investors. In the industry classification system, companies belonging to the same group should be similar, and dissimilar otherwise.

Because the market is driven, in part, by investor perception, the similarity between companies can be determined by calculating the co-movement of stocks. If market participants consider a set of companies to be closely related, those companies may experience similar movements in their stock returns, given similar exposure to risk factors and shocks that influence their returns. The use of return correlation to measure the similarity between companies has been explored earlier by [82] and [51]. We define the return correlation between the companies as a measurement of similarity.

For two stocks, i and j, at the end of year t, the similarity is given by $\rho_{i,j}$ where ρ is the correlation between the daily returns of stocks in year $t+1$. To evaluate the goodness of the prediction of k peers of company i, we calculated the average future

observed empirically that our transformation method led to similarly sized clusters, with higher values of the average return correlation between companies belonging to the same cluster.

returns correlation($\bar{\rho}_i$) given by: [8]

$$\bar{\rho}_i = \frac{\sum_{i \neq j} \rho_{i,j}}{k - 1}$$

After obtaining the average future returns correlation for each company, we obtain the goodness of the prediction as follows:

$$\bar{\rho} = \frac{\sum_{i=1,K} \bar{\rho}_i}{K}$$

where K represents the total number of stocks being evaluated, and $\bar{\rho}$ represents the overall average return correlation between similar stocks.

A higher value of $\bar{\rho}$ reflects better performance by the classification system in grouping similar stocks. In our evaluation, we used the $\bar{\rho}$ to assess the industry grouping models. While evaluating $\bar{\rho}$, we set values of $k = 2, 5, 10$, and also to 'dynamic'. Setting k as *dynamic* implies that the number of peers is decided by the number of companies in the sector, industry or sub-industry cluster to which company i belongs.

We evaluated the ability of the ML models to predict the peers of every company in our dataset. To do this, we evaluated the forward return correlation with the top k peers as predicted by the model. To compare against GICS, we decided the value of $k = \textit{dynamic}$ for a company by the number of peers in the GICS grouping to which it belongs. For example, at the industry level, if company i belongs to a GICS group that contains m peers, then $k = \textit{dynamic}$ at the industry level will include m peers in the calculation of $\bar{\rho}_i$. We evaluated peers at three levels: sector, industry and sub-industry. We evaluated the models for the years 2014-2020, and report the average values over all years in the study period.

Table 2.2 shows the effectiveness of the ML models in predicting the top k peers of each company. The ridge regression model was evaluated after being trained with features from individual datasets. We observed that the ridge regression model with

[8] The first peer of a firm i is firm i itself, hence we don't include that in the numerator and denominator.

all features (Ridge:ALL) outperformed the ridge regression model trained on only a subset of features. Adding GICS features to the model (Ridge:ALL+GICS) led to small improvements, which implies GICS features have information not captured by other data sources. We see will see later that adding GICS features also increased cluster stability. We also used nonlinear ML models, such as neural networks and XGBoost. In comparison to the ridge regression model, the nonlinear models showed greater peer-prediction power. We also include the performance of different ML models over each year individually in the Appendix in Table A.6, and we observed that ML models consistently outperformed GICS at the sector, industry, and sub-industry levels. In Table A.7 in the Appendix, we also include the standard errors for the forward returns correlation, using bootstrapping training data for the ridge regression model. We observed that standard errors were smaller than forward return correlation improvements by orders of magnitude. Hence, these improvements were statistically significant.

Model:Features	Sector	Industry	Sub-industry	Top 10	Top 5	Top 2
GICS	33.17	41.97	44.57	-	-	-
Ridge:Factors	33.69	39.99	42.13	40.80	42.16	44.15
Ridge:10k-ALL	33.62	40.66	43.50	41.54	43.38	46.02
Ridge:Returns	35.57	42.40	44.56	44.65	46.47	49.12
Ridge:ALL	36.20	43.75	46.22	46.16	48.35	51.55
Ridge:ALL+GICS	36.21	43.96	46.48	46.56	48.79	51.96
NN:ALL	36.13	43.64	46.09	46.30	48.48	51.65
NN:ALL+GICS	36.30	43.87	46.31	46.63	48.80	51.96
XGBoost:ALL	36.48	44.01	46.48	46.60	48.66	51.59
XGBoost:ALL+GICS	**36.58**	**44.26**	**46.69**	**46.84**	**48.95**	**51.96**

Table 2.2: Evaluation of ML models.
This table presents the values of average return correlation of peers in the out-of-sample periods. The peers are obtained via the ML models for different datasets. For a fair comparison with GICS, the number of peers for sector, industry, and sub-industry groupings was decided according to the number of companies in the GICS group to which the company belonged. ML models trained on all features outperformed GICS. However, the marginal performance improvements gained by adding GICS features (Ridge:ALL+GICS) and by the use of nonlinear ML models (NN:ALL, XGBoost:ALL) were small compared to ridge regression (Ridge:ALL). NN = neural network.

We have included the performance of peers obtained from the cosine distance metric in Table A.4 in the Appendix. On comparing cosine and ML models, we observed a minor improvement over GICS classification when using the cosine distance of features to find the top k peers. However, using the ML model's prediction as a similarity metric improved the return correlation between predicted peers by 2-3 percentage points over the GICS baseline and the best cosine model. This signifies that the ML models more effectively prioritized features that captured the market perception of companies than the cosine distance, which weights all features with equal importance. Thus, having unequal feature importance learned in a data-driven way led to improvements in performance.

After evaluating the peer-prediction power of the model, we next evaluated the clusters obtained after hierarchical clustering. From the predicted correlations, we computed the pairwise distance between companies and used that for hierarchical clustering, as described in the **Clustering** section. Table 2.3 presents the average return correlations for clusters at all three levels: sector, industry, and sub-industry. We determined the number of peers in the process of calculating average return correlation (the top k) based on the number of companies in the cluster.[9] We observed that ML clusters outperformed GICS by a large margin (up to 6 percentage points at the sub-industry level and 4 percentage points at the industry level). Adding GICS features as an input to the ML model also increased performance. The best-performing ML clusters were obtained from the ridge and XGBoost models trained with all features, including GICS. We include the performance distribution over all years after clustering for different models in Table A-1 in the Appendix. The improvements using ML models were consistent over the years. The clusters obtained from the different ML models were the different versions of AIPGS. In subsequent analysis, we observed that different versions had different desirable properties.

The clusters obtained from the ML methods grouped companies with similar market perception and higher return correlations. Apart from obtaining these groups, the ML models also provided a similarity score that can be used to obtain the top k peers

[9] Note that in Table 2.2, k was decided based on number of companies belonging to the same GICS group irrespective of model, while in Table 2.3, k was decided based on the number of companies belonging to the cluster from the corresponding model

Model:Features	Sector	Industry	Sub-industry
GICS	33.17	41.97	44.57
Ridge:ALL	32.04	44.10	48.75
Ridge:ALL+GICS	**33.54**	**44.73**	50.07
NN:ALL	30.70	41.09	48.20
NN:ALL+GICS	30.74	42.56	48.53
XGBoost:ALL	32.24	43.07	49.45
XGBoost:ALL+GICS	32.90	44.69	**50.64**

Table 2.3: Results after hierarchical clustering.
This table presents the average return correlation for clusters obtained by hierarchical clustering on the distance obtained from the ML models. Ridge:ALL refers to the ridge regression model with all the features from various datasets. +GICS refers to the addition of GICS features as one of the inputs to the model. The clusters at the industry and sub-industry levels had higher average return correlations between companies by up to 2.8 and 6 percentage points, respectively, compared to GICS clusters.

for any company. Next, we evaluated the different properties of clusters that are relevant to an industry classification system and measured their potential for diversification and hedging applications.

Similarity to GICS

Apart from the source of revenue, GICS also recognizes market perception as relevant for its classification purposes. Our aim is to create company groupings that incorporate market perception more effectively than GICS. Hence, we compared the similarity of our ML clusters to GICS groupings. We calculated this similarity using the adjusted Rand index (ARI), counting pairs of companies assigned to the same and different clusters in GICS groupings and ML clusters, normalized by the total pairs of companies.[10] Thus, the ARI represents the probability that a pair of companies will be grouped together by both GICS and our ML clusters. Higher ARI values imply greater similarity between the GICS groupings and ML clusters. Table 2.4 presents the ARI for different ML groupings at various levels. We observed that without using GICS input features, ML clusters were approximately 30%-50% similar

[10] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html

to GICS, and saw the greatest similarity at the industry level. Because GICS is a market-oriented classification system, some degree of similarity between the ML clusters and GICS is expected. Adding GICS features as an input increased the similarity to the GICS groupings; specifically, the similarity reached up to 71% for the ridge regression clusters.

Model:Features	Sector-ARI	Industry-ARI	Sub-industry-ARI
Ridge:ALL	32.81	48.45	44.91
Ridge:ALL+GICS	**58.59**	**71.10**	**63.12**
NN:ALL	26.15	39.18	39.09
NN:ALL+GICS	28.17	42.85	41.40
XGBoost:ALL	33.02	44.73	43.62
XGBoost:ALL+GICS	51.43	59.19	50.55

Table 2.4: Adjusted Rand Index for ML Clusters.
This table presents the ARI values for ML and GICS clusters. Ridge:ALL refers to the ridge regression model with all the features. +GICS refers to addition of GICS features as one of the inputs. We observed that ML clusters had similarities to GICS clusters—an expected result, given that GICS is a market-oriented classification model.

Granularity	GICS	Ridge	Ridge+GICS	NN	NN+GICS	XGB	XGB+GICS
Sector: 90%	97	35	24	41	31	30	26
Industry: 90%	97	54	40	24	29	20	32
Sub-industry: 90%	96	49	31	24	27	21	23
Sector: 50%	100	95	89	81	91	80	90
Industry: 50%	100	69	88	63	69	63	76
Sub-industry: 50%	99	62	75	56	60	54	65

Table 2.5: Cluster Stability.
In this table, we present the percentage of clusters retaining 90% and 50% of the companies over the years at different granularities of grouping. Ridge refers to the model trained on the all the features. Ridge+GICS refers to the ridge regression model trained on all features, including GICS features. We observed that up to 50% of clusters retained 90% of companies for ridge regression clusters, whereas up to 90% clusters retained 50% of companies for the majority of the models. Adding GICS features in different models generally increased the stability of the clusters. We concluded that, despite the dynamic nature of the clusters, the clusters were fairly stable, because the majority of them mapped from year t to year $t + 1$. GICS had the maximum stability. XGB = XGBoost.

Cluster Stability

We aim to create a dynamic classification system where the clusters update every year in tandem with the evolution of a company's businesses and market perception. However, cluster composition should not be drastically different from year t to year $t+1$; major changes in all the companies in our universe are unlikely, and stability is an important and desirable characteristic of a classification system. Hence, we analyzed the stability of clusters across different years. To compare their stability, we calculated the percentage of clusters that retain more than 90% and 50% of companies from year t to year $t+1$. We present the results in Table 2.5, in which we observed that up to 50% of clusters retained 90% of companies for ridge regression clusters, while up to 90% of clusters retained 50% of companies for a majority of the models. This showed that the ML clusters were not only dynamic, but fairly stable, and the majority of them were mapped from year t to year $t+1$. Adding GICS as one of the input features increased the stability of the clusters. The clusters from the ridge regression model were more stable than those from XGBoost or neural networks.We also observed that GICS grouping had the maximum stability which is a desirable property of a classification system.

Cluster Factor and Industry Effects

We expect that companies belonging to the same clusters will react similarly to different systematic factors. Hence, each cluster of companies can be considered as a factor that can explain the systematic shocks common to the group. We analyzed the clusters obtained from the ML model as factors and evaluated their homogeneity and out-of-sample diversification benefits . We followed the methodology of [142] to calculate the industry effects, which involves calculating the cluster factor returns using the following cross-sectional regression for every trading day t:

$$R_{j,t} = A(t) + \sum_{i=1}^{N} L_{i,t} D_{j,i,t} + \epsilon_{j,t}$$

where $R_{j,t}$ is the return of stock j at time t, $D_{j,i,t}$ is an indicator variable which is equal to 1 if stock j belongs to cluster i in time t, and $\epsilon_{j,t}$ is the return specific to stock j at time t. $L_{i,t}$ is the return of cluster i at time t, which is learned by the model. $A(t)$ is the intercept for the aforementioned cross-sectional regression. We performed a weighted regression using the square root of market capitalization to give more, but not excessive, weight to companies with high market capitalization, following the methodology of the MSCI Barra US Total Market Model.

Granularity	GICS	Ridge	Ridge+GICS	NN	NN+GICS	XGB	XGB+GICS
Sector	0.73	0.85	0.80	0.85	0.85	**0.87**	0.84
Industry	0.93	**1.15**	1.13	1.12	1.14	1.13	1.09
Sub-industry	1.21	1.54	**1.56**	1.39	1.43	1.36	1.38

Table 2.6: Diversification Table for Cluster Factor Returns.
This table presents the standard deviation of cluster factor returns (Equation 2.1) averaged over time for clusters at different levels of granularity. The cross-sectional dispersion of ML clusters was up to 30% higher than GICS clusters. A higher standard deviation for ML cluster returns implies greater diversification potential using cluster factors and greater power in explaining the cross section of stock returns.

Homogeneity of clusters in terms of return characteristics is verified using forward return correlation. Here, we seek to evaluate the diversification benefits of ML cluster factors. We calculated the cross-sectional dispersion (the standard deviation) for cluster factor returns, as suggested by [262]. A higher cross-sectional dispersion implies greater power in explaining the cross section of security returns, which may help investors as they seek to build better and more efficient portfolios when used in the context of a factor model and optimizer.

$$D_t = std_i(L_{i,t}) \tag{2.1}$$

We averaged D_t and present the results in Table 2.6. The cross-sectional dispersion of ML clusters was up to 30% higher than that of GICS clusters. Hence, the ML clusters showed a potential benefit in portfolio construction over GICS.[11]

[11] Although portfolios built using ML clusters may incur higher turnover costs owing to their lower stability compared to GICS.

To analyze the ability of ML models to group companies with similar fundamentals, we also evaluated the fundamental characteristics of companies grouped together as measured by their size, leverage, profitability and so on. We calculated the fundamental factor exposures for different clusters by averaging the factor exposures of companies belonging to the cluster and calculated the diversification benefits for each fundamental factor using the cross-sectional standard deviation of a cluster's factor exposure, similar to Equation 2.1. Along with the inter-cluster dispersion of fundamental factor exposures, we calculated the intra-cluster dispersion of exposures by calculating the standard deviation of factor exposures of companies in a cluster and averaging the standard deviations of all clusters. Smaller values of intra-cluster dispersion signify that companies belonging to the same cluster have similar fundamental values.

Tables 2.7 and 2.8 present the intra- and inter-cluster dispersion for market beta, dividend yield, earnings yield, growth, size, and residual volatility for clusters at different levels. For factors such as residual volatility, dividend yield, and size, the intra-cluster dispersion was smaller than it was for GICS clusters. Similarly, the inter-cluster variances of factor exposures for residual volatility, size, and earnings yield were 40% higher than they were for the GICS clusters. Lower values of intra-cluster dispersion for ML clusters imply that the ML models grouped together companies with similar fundamentals, and higher values of inter-cluster dispersion imply that ML clusters have greater diversification potential for portfolios. Because fundamental factors are one of the inputs to the ML models, the models learned to group together companies with similar fundamentals. Hence, along with grouping companies with similar return characteristics, the ML clusters also grouped companies with similar fundamental factor exposures.

Hedging Risk Exposure

Prior to clustering, we obtain a list of similar stocks, also known as peers, for each company, using both the cosine and the ML models. These peers have similar risk exposures and return characteristics because they have higher values of future returns

Granularity	GICS	Ridge	Ridge+GICS	NN	NN+GICS	XGB	XGB+GICS
			Market Beta				
Sector	0.82	0.79	0.79	0.76	**0.76**	0.76	0.78
Industry	0.70	0.72	0.72	**0.69**	0.70	0.70	0.70
Sub-industry	**0.65**	0.67	0.66	0.66	0.66	0.66	0.66
			Dividend Yield				
Sector	0.84	0.75	0.76	0.75	0.74	**0.72**	0.75
Industry	0.71	0.68	0.68	**0.66**	0.66	0.67	0.68
Sub-industry	0.70	0.62	0.62	0.60	0.60	0.60	**0.59**
			Earnings Yield				
Sector	0.81	0.79	0.79	0.78	0.76	**0.76**	0.77
Industry	**0.68**	0.74	0.72	0.68	0.69	0.69	0.70
Sub-industry	**0.63**	0.68	0.68	0.66	0.67	0.67	0.67
			Growth				
Sector	0.91	0.93	0.93	0.91	**0.90**	0.92	0.93
Industry	**0.79**	0.88	0.86	0.85	0.86	0.86	0.87
Sub-industry	**0.75**	0.81	0.82	0.82	0.82	0.83	0.83
			Residual Volatility				
Sector	0.79	0.76	0.77	0.71	0.72	**0.71**	0.74
Industry	0.71	0.73	0.72	**0.66**	0.67	0.67	0.68
Sub-industry	0.66	0.68	0.67	0.64	0.65	**0.64**	0.65
			Size				
Sector	0.98	0.94	0.96	**0.90**	0.91	0.90	0.92
Industry	0.93	0.81	0.87	0.78	0.79	**0.78**	0.81
Sub-industry	0.88	0.71	0.75	0.68	0.71	**0.66**	0.68

Table 2.7: Intra-Cluster Dispersion for Various Fundamental Factors.
The above table presents the intra-cluster dispersion of factor exposures for various clusters. Different columns refer to different ML methods. For example, Ridge refers to the clusters obtained using ridge regression on all features, excluding GICS. Ridge+GICS refers to the ridge regression model trained on all features, including GICS features. For factors such as size, residual volatility and dividend yield, the intra-cluster dispersion for ML clusters was smaller than GICS clusters at all levels. Lower values of intra-cluster dispersion imply that companies in clusters have similar fundamental characteristics.

Granularity	GICS	Ridge	Ridge+GICS	NN	NN+GICS	XGB	XGB+GICS
			Market Beta				
Sector	0.69	0.77	0.74	**0.81**	0.77	0.80	0.76
Industry	0.68	0.69	0.70	**0.73**	0.74	0.73	0.72
Sub-industry	0.72	0.74	0.75	0.75	**0.76**	0.75	0.75
			Dividend Yield				
Sector	0.60	0.73	0.68	**0.76**	0.75	0.76	0.73
Industry	0.66	0.66	0.65	**0.73**	0.72	0.73	0.70
Sub-industry	**0.73**	0.69	0.68	0.71	0.71	0.72	0.71
			Earnings Yield				
Sector	0.44	0.66	0.62	0.69	0.66	**0.71**	0.67
Industry	0.50	0.67	0.67	0.68	**0.70**	0.69	0.67
Sub-industry	0.57	0.78	0.78	0.78	0.78	**0.79**	0.78
			Growth				
Sector	0.34	**0.41**	0.38	0.40	0.38	0.40	0.39
Industry	0.45	0.48	0.46	**0.50**	0.48	0.47	0.46
Sub-industry	0.53	**0.63**	0.63	0.61	0.60	0.57	0.57
			Residual Volatility				
Sector	0.49	0.66	0.63	0.72	0.72	**0.73**	0.69
Industry	0.53	0.68	0.68	0.70	**0.71**	0.70	0.69
Sub-industry	0.61	0.78	**0.79**	0.78	0.79	0.78	0.79
			Size				
Sector	0.22	0.32	0.23	**0.37**	0.34	0.36	0.30
Industry	0.44	0.58	0.49	0.62	0.58	**0.63**	0.57
Sub-industry	0.62	0.70	0.68	0.73	0.69	**0.73**	0.72

Table 2.8: Diversification Table.

This table presents the standard deviation of a cluster's fundamental value averaged over time for clusters at different granularity. Different columns refer to different ML methods. For example, Ridge refers to the ridge regression model trained on all features, excluding GICS. Ridge+GICS refers to ridge regression trained on all features, including GICS. For factors such as residual volatility, size, and earnings yield, the cross-sectional standard deviation was around 40% higher than GICS clusters. Higher standard deviation of fundamental factor exposures for ML clusters implies higher cross-sectional dispersion in cluster factors.

correlations. Hence, the peers can be used for hedging a portfolio's risk exposure. We demonstrate this by constructing single-stock portfolios. For every company i in our universe, we constructed a U.S.-dollar-neutral portfolio by taking a long position on company i and an equally weighted short position on its top k peers. We determined the top k peers using different peer-prediction models, which were either the cosine or ML model. For this analysis, we used ML models that were trained on features calculated from all datasets combined. We performed a comparison of different models with their GICS peers. From the GICS grouping, we obtained a set of companies belonging to the same GICS industry. From the same industry cluster, we selected the top k peers on the basis of size, as done in the literature [9]. The above methodology is inspired by [296]

After constructing these U.S.-dollar-neutral portfolios, we calculated their daily returns and the value of their annualized volatility. We then averaged the realized volatility for all single-stock portfolios, as shown in Table 2.9. If the peers have similar risk exposures and return co-movements, we expect the realized volatility of the hedged portfolios to be smaller. Based on Table 2.9, using one and five peers for hedging single-stock portfolios decreased the realized volatility by 643 and 259 basis points, respectively, when compared to size-based peers obtained from GICS groups.

Model	GICS	Ridge	Ridge+GICS	NN	NN+GICS	XGB	XGB+GICS	Change
$k = 1$	43.88	38.88	38.23	37.81	37.67	37.49	**37.45**	643
$k = 5$	34.21	32.02	31.91	31.90	31.81	31.70	**31.62**	259

Table 2.9: Hedging Risk Exposures.
We present the average annual realized volatility (as percentages) of hedged single-stock portfolios when peers were decided based on different models. The rows correspond to the number of companies selected for hedging. The Change column refers to the difference in the realized volatility of GICS and the best model (XGB+GICS).The volatility was reduced by up to 643 and 259 basis points for $k = 1$ and $k = 5$, repectively, when peers were obtained from the XGBoost model.

Model Interpretation

We have evaluated the abilities of different ML models against GICS groupings to predict the similarity of companies and cluster them effectively. Here, we interpret the predictions of these ML models and find the top-contributing features and datasets that captured the similarity between companies.

We observed that XGBoost was the best-performing model for various metrics. It outperformed ridge regression by a small margin in the average return correlation. For ridge regression, calculating the contribution of features was trivial, given the linearity of the model. For XGBoost, we used the post-hoc method of interpretability, TreeSHAP [194], which calculates the Shapley values of the features that contributes to the prediction of the output.

We present some examples of predicted peers and the contribution of their features. Figure 2-2 presents the top-10 peers for eBay predicted for the year 2014 by the XGBoost model. We include the predicted correlation and the actual correlation between eBay and its peers. The first six columns show the contributions of the different dataset features. We have highlighted two peers, The American Express Company (AmEx) and IBM. Even though the GICS contribution for AmEx was negative, the XGBoost model determined AmEx to be a peer due to its high similarity in factor exposures and 10-K filings. We observed that the actual correlation between AmEx and eBay was high in 2014. In the case of IBM, the actual correlation was not high. However, there appears to have been significant contribution by news co-occurrence features. This implies that IBM was mentioned with eBay multiple times in news articles, which led to it being predicted to be one of eBay's peers. On further analysis, we found that the news articles mentioning eBay and IBM together were earnings announcements on the same dates and did not convey much information about the similarity between the companies. Similarly, we can analyze the feature contributions for individual predictions of all companies by using TreeSHAP for both XGBoost and the model coefficients for ridge regression.

Next, we calculate the global-feature importance over the years for the ridge re-

				Ebay 2013				
RET	GICS	Factor	10k-TFIDF	10k-D2V	NEWS	Top-10 Peers	PRED CORR	CORR
0.016	-0.004	0.073	0.007	0.000	0.005	DANAHER CORP DEL	0.436	0.288
-0.021	0.041	0.039	0.009	0.014	0.014	YAHOO INC	0.437	0.383
0.001	0.026	0.050	0.017	0.003	0.000	TOTAL SYS SVCS INC	0.437	0.374
0.009	0.019	0.054	0.000	-0.005	0.021	INTERNATIONAL BUSI	0.438	0.191
-0.025	0.052	0.067	0.014	-0.006	0.000	J2 GLOBAL COMMUNI	0.442	0.338
-0.008	-0.004	0.058	0.053	-0.002	0.016	AMERICAN EXPRESS C	0.454	0.425
-0.003	0.020	0.057	0.038	-0.007	0.020	VISA INC	0.466	0.341
0.000	0.033	0.043	0.042	0.011	0.002	VANTIV INC	0.471	0.236
0.019	0.030	0.081	0.013	-0.004	0.001	FIDELITY NATL INFORM	0.480	0.383
0.011	0.030	0.041	0.046	0.001	0.022	MASTERCARD INC	0.492	0.346

Figure 2-2: Predictions of eBay Peers for 2014.
The table presents the contribution of features from different datasets for the prediction of the top-10 peers of eBay for the year 2014. Features with a high contribution are colored with a stronger degree of red, while features with a weaker contribution are colored by blue. The high similarity between AmEx and eBay can be attributed to the factor exposure and 10-K-TFIDF datasets, even though they belonged to different GICS groups. For the predicted correlation (PRED CORR) and observed correlation (CORR), we colored AmEx green, because it is shown to be the most similar to eBay, and colored IBM red, because it had the least actual similarity, despite being seen as one of eBay's top-10 peers. D2V=Doc2Vec.

gression model. We aggregated the feature importance (i.e., the absolute value of a coefficient) over different years, and present the top 10 contributing features in Figure 2-3. We observed that the past returns correlation was the most significant feature, followed by the GICS sub-industry and industry features. The ridge regression model gave maximum importance to the return correlation from historical data, suggesting that the market perception (return correlation) of companies did not drastically change over the years. GICS features were also some of the top contributing features, which was expected, because GICS is a market-oriented classification system, and we are aiming to use market perception to group companies. Feature importance, as presented in Figure 2-3, shows the variables that had the largest impact on the predicted forward return correlation.

We also grouped individual features by their datasets, and present the overall feature importance in Figure 2-4. We observe that the returns dataset had the highest contribution, followed by factor exposure and 10-K information.

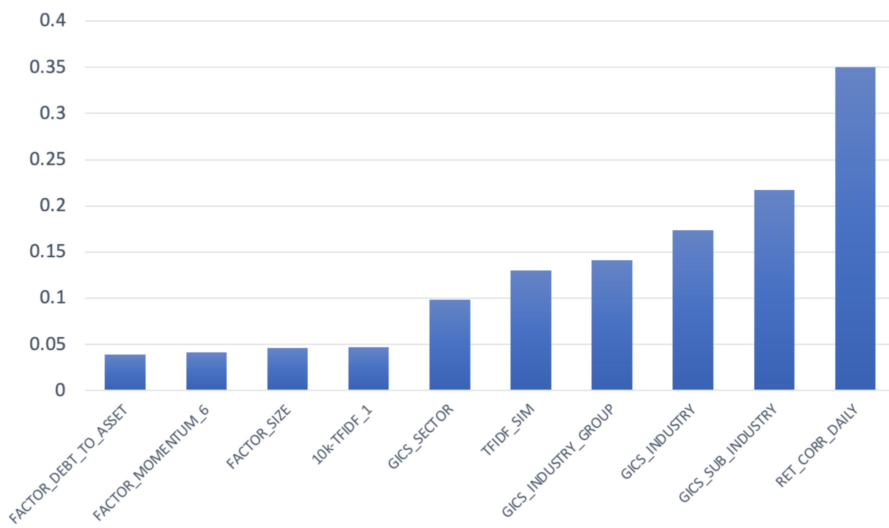


Figure 2-3: Feature Importance.
The plot shows the 10 most significant features for the ridge regression model aggregated over testing years. The return correlation as calculated from historical data was most important, followed by certain GICS features. Some fundamental factors such as size, momentum, and the debt-to-asset ratio were also found in the 10 most important features.

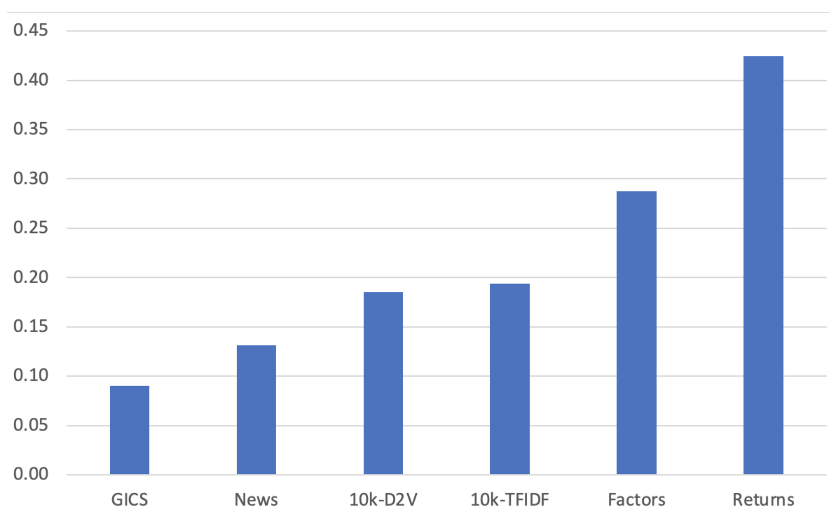


Figure 2-4: Feature Importance Aggregated by Dataset.
The plot shows the importance of features obtained by the absolute value of coefficients from different datasets and aggregated over years for the ridge regression model. The returns dataset was most important, followed by factor exposure and 10-K information.

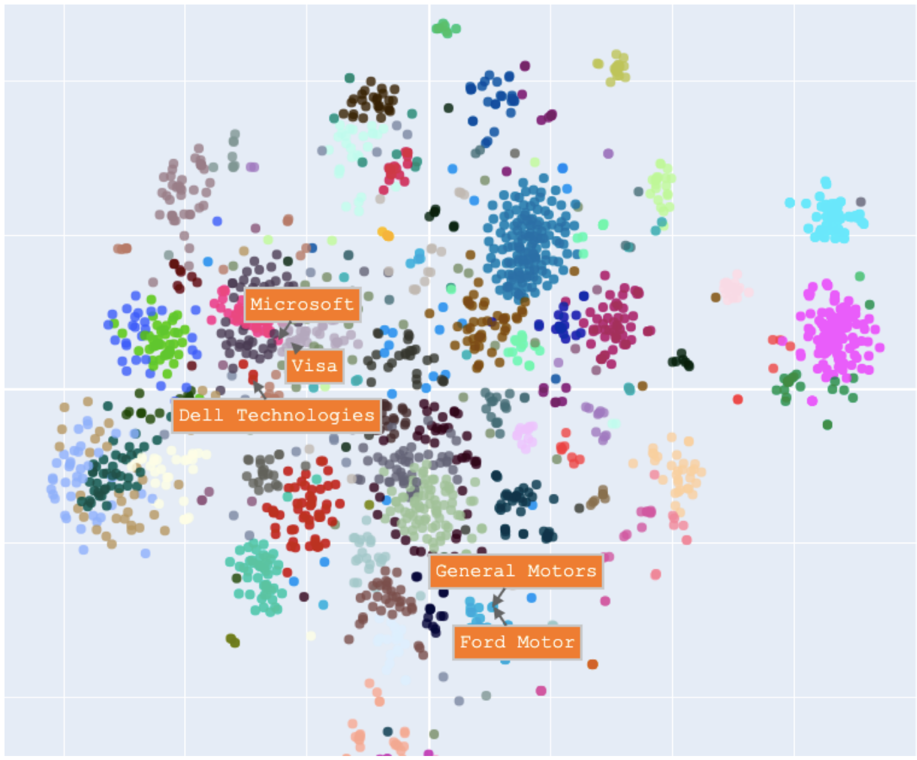
Cluster Visualization

In our sample, we grouped approximately 2,000 companies into a different number of clusters every year, depending on their sector, industry, or sub-industry level. This makes analyzing and visualizing peers of these companies and classification a non-trivial task. Hence, we developed a cluster visualization system with multiple functionalities; this allowed us to visualize the companies and their peers in the cluster groups and their nearest neighbors.

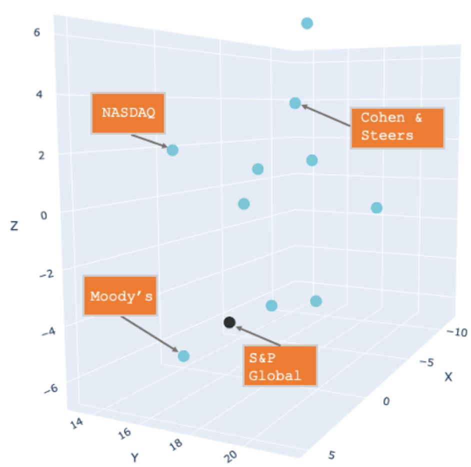
Given the distance matrix D, $D_{i,j}$ is the distance between companies i and j as predicted by the ML models. We aim to learn a 2D or 3D vector X_i and X_j for companies i and j, which approximates $distance(X_i, X_j)$ as $D_{i,j}$. To do this, we used t-distributed stochastic neighbor embedding (t-SNE) [278]. This constructs a probability distribution over pairs of companies in high- and low-dimensional space. It learns the lower dimensional coordinates by minimizing the KL divergence between the probability distributions in low- and high-dimensional space. To visualize the clusters, the lower-dimensional vector X, obtained from t-SNE, can be plotted for the companies.

The visualization system contains three modes. In the first mode, "Overall," we plotted all the companies on a 2D plane. The color for each point is assigned based on the cluster to which it belongs. For example, in Figure 2-5a, we present all of the companies for the year 2020 and assign colors based on the ML model's industry-level clusters. We can observe that General Motors Company and Ford Motor Company, both being automobile firms, belonged to the same industry group. Similarly, Dell Technologies (technology, hardware, storage), Visa (IT services), and Microsoft Corporation (software) were plotted close to each other, as they were learned to be similar by our model.

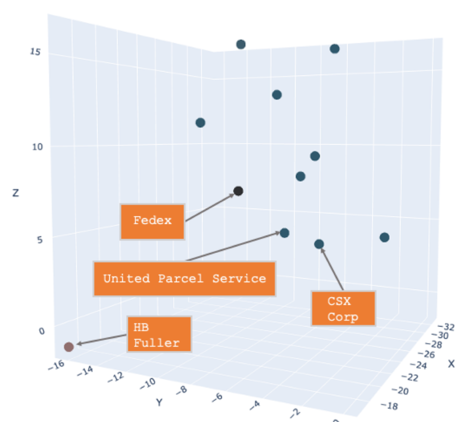
In the second mode, "Kernel(s)," we can select at least one company and plot the other companies belonging to the same cluster group on the 2D/3D plane. Figure 2-5b shows one such example for S&P Global's ML sub-industry clustering for the year 2020. We can see that the NASDAQ Stock Market and Moody's Investor Service were assigned to the same sub-industry group; they belong to the same GICS sub-industry group (financial exchange and data) group as S&P Global, whereas Cohen & Steers Inc., which belongs to “asset management and custody bank" in GICS, is also assigned to the same ML sub-industry. In the third mode, “Nearest-Neighbor," we can plot up to the 20 most-similar companies for a selected company X; these companies do not necessarily need to be classified in the same group as company X. Figure 2-5c shows the 10 nearest neighbors for FedEx Corporation (air freight and logistics). As expected, other transport companies like United Parcel Services (air freight and logistics) and CSX (railroads) were found to be similar; interestingly, H.B. Fuller Company (chemical), was also determined to be one of its 10 nearest neighbors (a chemical company according to GICS). These visualization frameworks can be accessed via https://dash-app-vis.herokuapp.com/.



(a) Overall mode. This figure shows all the companies for 2020 with a color assignment based on the industry group determined by the ML model. We observed that similar companies like Ford and General Motors belonged to the same industry clusters. Similar companies such as Microsoft, Visa, and Dell Technologies were placed close to each other as well.



(b) Kernel(s) mode. This figure shows the ML sub-industry cluster distribution for S&P Global for 2020.



(c) Nearest-neighbor mode. This figure shows the 10 nearest neighbors for FedEx in 2020.

Figure 2-5: Examples from the Cluster Visualization Framework.

2.6 Conclusion

In this chapter, we developed an AI-driven peer grouping system, AIPGS, to group stocks according to market perception. These industry groupings are dynamic and updated as the characteristics of companies in the market evolve over time. As evidenced by its higher return correlations, AIPGS outperformed the closest benchmark, GICS, in grouping stocks. AIPGS uses information from multiple different data sources, including returns, factor exposure, 10-K filings, and news, to learn the similarity between companies, and uses that similarity to create groups of homogeneous companies.

AIPGS clusters had high intra-cluster similarity and high inter-cluster dispersion for financial ratios and fundamentals. We verified this for different fundamental characteristics including market beta, dividend yield, earnings yield, growth, residual volatility and size. These data-driven clusters grouped companies with similar fundamental characteristics, and we saw a higher level of diversification potential for portfolios constructed using ML clusters.

Apart from grouping stocks, our models predict a score for every pair of firms in our universe to reflect their future returns correlation (similarity). This can be used to obtain the nearest neighbors of a company, and to help investors understand which companies appear to be the most similar according to the market. In this work, we showed one use case of similarity scores: hedging single-stock portfolios.

This work showed that a classification system could be obtained using AI tools that performed better than traditional classification systems in some metrics, but the use of AI tools creates an interpretability problem. Although we can quantify the contributions of the datasets and various features for the similarity score of a pair of peer companies, we cannot assign a summary to the ML clusters that explains why the companies were grouped together. Thus, for some use cases where a premium is placed on interpretability and stability, traditional classification systems may be preferred. In other situations, where estimating future returns correlation is most important, our framework may dominate.

Regardless, we believe the framework presented here could be a valuable tool for investors, with potential applications for portfolio construction, hedging, and other areas. As an extension of this work, the ML clusters can be combined with the insights of experts and analysts to assign a summary to the clusters that can be interpreted by humans, as exists for GICS clusters. We also believe that an emphasis on greater interpretability may increase the accessibility of the ML clusters.

As we developed different versions (clusters from different ML methods) of AIPGS and compared them with GICS, we expect that different systems could be used, depending on the requirements of the user. We were able to obtain the best peers and clusters (in terms of forward returns correlation and similar risk–return profile) using the XGBoost model and sacrificing interpretability and stability of clusters. GICS had the most stable and interpretable clusters. Clusters and peers from the ridge regression model had a fair balance of stability and performance.

For investors, our peer grouping system, in conjunction with other existing classification systems, has the potential to be very insightful in investment diligence. For example, an investor looking for exposure to insurance brokers might include companies present in the GICS "insurance brokers" sub-industry. However, our model suggested including Berkshire Hathaway Inc. in this group. More specifically, for the year 2019, the ML model predicted that Berkshire Hathaway, a major conglomerate with investments across multiple industries, should be grouped in the same sub-industry as other firms that GICS would consider insurance brokers. On further analysis, this classification is meaningful because the market may perceive Berkshire Hathaway as an insurance broker, due to the fact that one of its most prominent companies, The Government Employees Insurance Company (GEICO), is an insurance underwriter. Because AIPGS provides additional insights beyond GICS, it can used as an extension of GICS. Furthermore, the similarity scores and clusters produced by our model may be used by some for portfolio management, such as investors seeking to improve hedging techniques.

Apart from the transparency and stability of ML-based clusters, another limitation of our approach is that our ML models can only group public companies owing to unavailability of data on private companies. Extending the classification system to non-public companies is one important extension that we hope to pursue in the future. Lastly, we can try to include other data sources, including analyst-coverage overlap, supply chain overlap, and broker research, that might contain extra information not captured in the datasets used in this work.

Chapter 3

Peer Grouping: Crypto-Connections:- Exploring the Peers

In this chapter, we presents a novel data-driven approach for developing a cryptocurrency peer grouping system and similarity score measure for application in investment management. We collected data from multiple sources, including returns, fundamental factors, and whitepapers, to construct our similarity scores. Our analysis demonstrates that similarity scores is a significant predictor of future return correlation between pairs of coins, and can also be utilized to reduce the volatility of single currency portfolios. Furthermore, we highlight the appealing risk-return characteristics of the Bitcoin cluster obtained from our peer grouping system.

3.1 Introduction

Cryptocurrency, a digital asset built on the foundation of cryptography and decentralized technologies, has garnered significant attention in recent years for its potential to revolutionize traditional financial systems. The development of the first cryptocurrency can be traced back to the creation of Bitcoin, which was introduced in 2008 by an individual or group of individuals using the pseudonym Satoshi Nakamoto. The release of the Bitcoin whitepaper, titled "Bitcoin: A Peer-to-Peer Electronic Cash System," laid the foundation for the cryptocurrency revolution. Since then many more cryptocurrencies have emerged and right now there are as many as 10000+ [1].

Cryptocurrency has emerged as a novel and disruptive asset class, capturing the attention of investors seeking to diversify their portfolios and tap into the potential of decentralized digital assets. The global market capitalization of cryptocurrency market is 1.3 trillion USD[1]. The investment in crypto-currrency allows investors to diversify their portfolio, as they tend to have low correlations with traditional asset classes like stocks, bond etc. Along with the diversification, there are potential of high returns over years. For example, $1 invested on January 1, 2016 in Bitcoin would yield a return of 7,478% to date[2]. It also allows investors to participate in new technologies such as decentralized finance (DeFi) which could lead to further value creation in the long run. Hence, the study of cryptocurrencies for investment is essential.

One of the tools necessary for portfolio management is the peer grouping system. For example, there exists a global industrial classification system (GICS) for industries and data-driven classification systems proposed in [35]. These systems can be used for obtaining peers which can be used for portfolio hedging, risk management, etc.

Similar to the industries, the peer grouping tools are necessary for cryptocurrencies investment management. There exist cryptocurrency taxonomies and they are not developed with the investment management application. They are based on market capitalization [289], architectural features [118], technical/financial/business [159], consensus system [126] and few more [165, 270].

The existing peer grouping system for industries (like GICS, data-driven systems, [35]) that captures market perception. Market perception is defined as the similarity in the returns between the companies. It is important because it allows stakeholders to compare peers based on their financial performance, which is important for investment management.

In this work, we propose a data-driven peer grouping method for the cryptocurrencies from 2016-2021. We use different datasets:- returns, price-volume and

[1] Information obtained from CoinGecko on April 13, 2023
[2] April 13, 2023

whitepapers and use these datasets to develop similarity scores and clusters of similar crypto-currencies that are focused with application in investment management. We conduct multiple tests to prove that similarity scores are predictive of the the future similarity (future return correlation) of coins. We hedge the single-stock portfolios and show that peers obtained from multiple similarity score versions reduces the volatility of the single stock portfolios. We also construct portfolios from the cluster of the crypto-currencies and present the return characteristics of the cluster portfolios.

The similarity score and the return co-movement for the pair of crypto-currencies is not extensively studied in the existing literature. One of the closest study is investigation of the code similarity and correlation of crypto-currency returns [192]. Our work can incorporate the code/developer similarity in the computation of similarity score with the availability of relevant data. [245] studied the co-movement of the four coins and [248] studied the pairwise correlation of many crypto-currencies and presented the interactions between currencies as the networks. [202] studied the properties of crypto-currencies and [201] showed that prices of bitcoin are correlated to the trading activities of the bitcoin owners with large ownership. Our study is different from existing literature in multiple aspect. First, we create the similarity score measure using multiple datasets and show that similarity score measures are indicative of future return correlation. Second, we cluster the similar coins (which can be considered as risk factors) and compute the return characteristics of the clusters.

Next, in Section 3.2, we present the details of the datasets and data processing steps. Section 3.3 presents the different tests we perform to show the utility of the similarity score and conclude in the Section 3.4.

3.2 Methodology

For this study, we collected multiple datasets: whitepapers and price-volume. In this section, we describe how we created these datasets and outline the various preprocessing steps.

3.2.1 Data

Whitepapers

Whitepapers for crypto-currencies are the documents that describe the design, purpose and underlying technology for the particular cryptocurrency. They are written by the founder or designer and the purpose of whitepaper is to provide detailed explanation of the working of project.

Whitepapers typically include the following: objectives, technical architecture and design, consensus mechanism, token distribution and economic model, governance and decision-making process, security features and potential risks, use cases and applications, roadmap and development plan, amongst other information. Since the whitepapers include significant important information about crypto currencies, we use it as one of the source for developing a similarity score and classification system.

One of the drawbacks of using whitepapers as a data source for classification is that not all cryptocurrencies have whitepapers. Therefore, in our study, we utilize the available whitepapers for the cryptocurrencies. In Figure 3-1, we display the number of coins with accessible whitepaper data. In total, we have whitepapers for 99 coins. The dataset was obtained from `https://www.allcryptowhitepapers.com/` and was converted into text format using optical character recognition from the PDFs.

Returns and Market Capitalization

We collect monthly returns of cryptocurrencies for this study. The data was scraped from the CoinGecko API. CoinGecko computes the prices of cryptocurrencies using the pair trading data from different exchanges. It aggregates prices from multiple crypto exchanges using an aggregation algorithm. Figure 3-2 displays the number of coins with price-volume data in our dataset. We observe that the number of coins has increased significantly over time. Along with returns, we also scraped market capitalization data from the CoinGecko API. In the next section, we discuss how we utilize the price-volume and market capitalization data for computing fundamental factors for cryptocurrencies. We utilized the CCI30 as a crypto market index. It

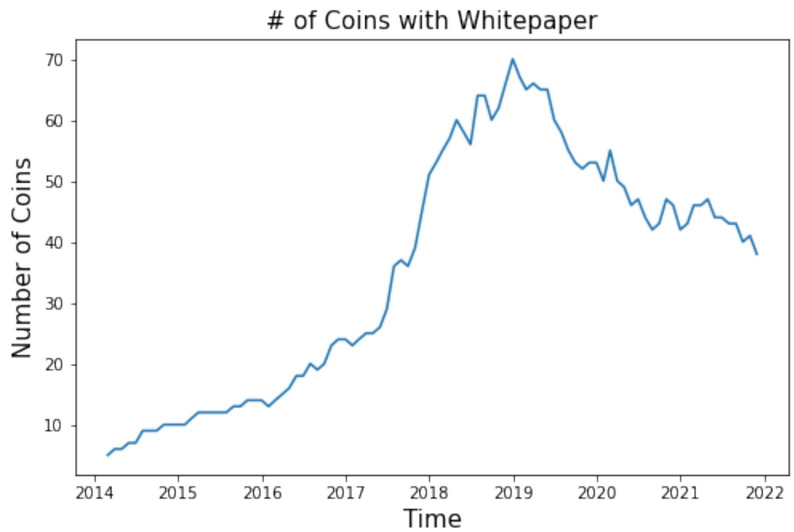


Figure 3-1: Number of Coins with Whitepapers in our Dataset.

tracks the overall crypto market using the top 30 coins by market capitalization. The data for the CCI30 market index is available at *cci30.com*.

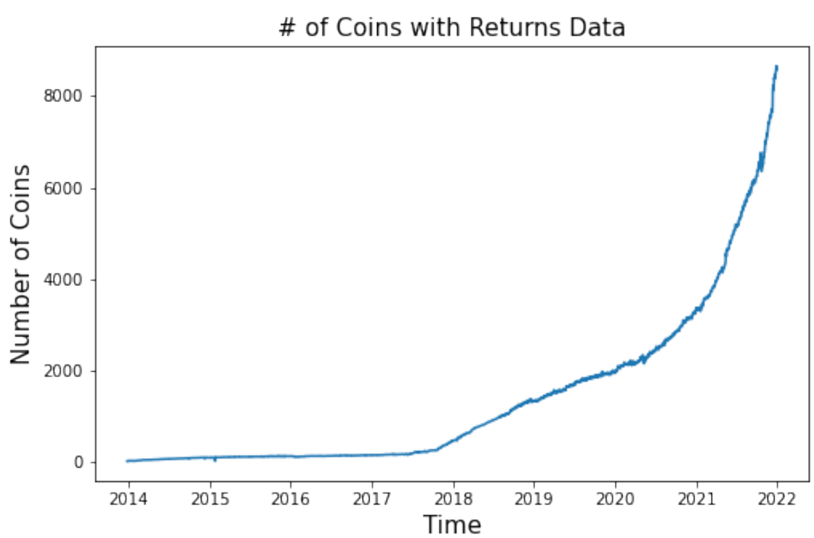


Figure 3-2: Number of Coins with Returns in our dataset.

Figure 3-3: Number of Coins in top x%age of market capitalization.

3.2.2 Featurization

White-papers

Whitepapers consist of textual data; therefore, we employed natural language processing to extract relevant features. The first step in pre-processing includes removing stop words, punctuation, numbers, etc. The next step is lemmatization, which involves reducing words to their base form. Finally, we extract features using a method known as TF-IDF, which stands for Term Frequency-Inverse Document Frequency.

Fundamental Factors

There is an extensive asset pricing literature on equities that includes a variety of factors and factor models developed to explain the cross-section of equity returns [106, 122].

Although the literature on fundamental factors for crypto-currencies is still emerging, some existing research has explored the factors that explain the cross-section of returns [178, 175, 176]. As exposure to these factors is a key driver of returns in crypto-currencies, we develop similarity scores using fundamental factors.

Following the existing literature, the fundamental factors presented in Table 3.1

were computed for each coin using data from 2015 to 2021. For each factor, we used the lookback $window = 30, 180, 365$ days, with the exception of factor RET, where we used $window = 1, 2, 3, 4, 8, 16, 50, 100$ weeks. Each coin i at time t was represented by the vector $\vec{F_{i,t}}$ of factor values $f_{i,t}$.

Factor Name	Factor Calculation
MCAP	$\texttt{MCAP}_t = \log(\texttt{market cap}_t)$
PRC	$\texttt{PRC}_t = \log(\texttt{price}_t)$
MAXDPRC	$\texttt{MAXDPRC}_t = \max(\texttt{prices}_{t-window+1:t})$
RET	$\texttt{RET}_t = \prod_{i=t-window+1}^{t}(1 + \texttt{return}_i) - 1$
VOL	$\texttt{VOL}_t = \log(\texttt{volume}_t)$
PRCVOL	$\texttt{PRCVOL}_t = \log(\texttt{price}_t) \cdot \log(\texttt{volume}_t)$
VOLSCALED	$\texttt{VOLSCALED}_t = \log(\texttt{market cap}_t) \cdot \log(\texttt{price}_t) \cdot \log(\texttt{volume}_t)$
BETA	$\texttt{BETA}_t = \beta$ such that $\texttt{returns}_{t-window+1:t,coin} = \beta \cdot \texttt{returns}_{t-window+1:t,market} + \alpha$, where $\texttt{returns}_{t-window+1:t,coin}$ are the daily returns for a coin and $\texttt{returns}_{t-window+1:t,market}$ are the daily returns for the CCI30 Market Index
BETA2	$\texttt{BETA2}_t = (\texttt{BETA}_t)^2$
IDIOVOL	$\texttt{IDIOVOL}_t =$ standard deviation of the residual ϵ after estimating $\texttt{returns}_{t,coin} = \beta \cdot \texttt{returns}_{t,market} + \alpha + \epsilon$
RETVOL	$\texttt{RETVOL}_t =$ standard deviation $(\texttt{returns}_{t-window+1:t})$
MAXRET	$\texttt{MAXRET}_t = \max(\texttt{returns}_{t-window+1:t})$
STDPRCVOL	$\texttt{STDPRCVOL}_t =$ standard deviation $(\texttt{PRCVOL}_{t-window+1:t})$
DAMIHUD	$\texttt{DAMIHUD}_t = (\frac{1}{window}\sum_{i=t-window+1}^{t} \texttt{returns}_i)/\texttt{PRCVOL}_t$

Table 3.1: The list of fundamental factors computed for each crypto-currency coin.

3.2.3 Hierarchical Clustering

Given the similarity score, next, we construct the clusters (groups) of similar crypto-currencies. Taking inspiration from [35, 204, 76], we construct the clusters using hierarchical clustering algorithms. The clustering should have 1) homogeneity:- similar currencies belong to the same cluster and 2) fair balance:- the clusters size should be fairly uniform. Hence, we use agglomerative (bottom-up) clustering with ward linkage. Ward linkage minimizes the intra-cluster variance when merging clusters which ensures that cluster size is fairly uniform. The number of clusters in the hierarchical clustering can be selected on the basis of the application (low: if big clusters are desired and high: if small clusters are desired). Once we have similarity scores and clusters of similar currencies, we analyze them in next Section.

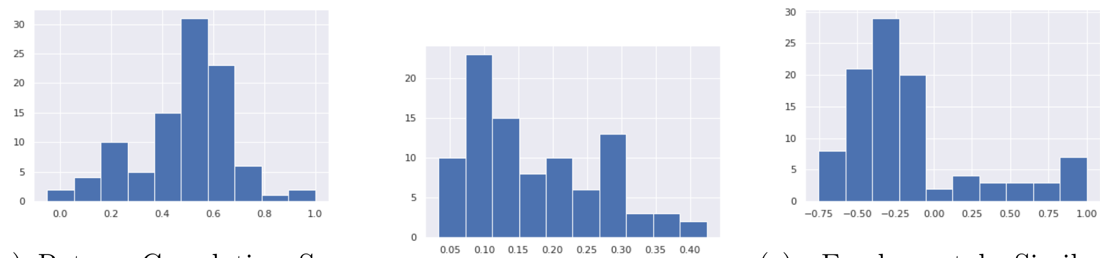
3.3 Results

In the previous section, we defined three distinct approaches for calculating similarity scores: whitepaper similarity, fundamental factors similarity, and returns similarity. In this section, we will evaluate the effectiveness of each method in computing similarity.

Since similarity scores are based on either correlation or cosine distance, they range from -1 to 1. Figure 3-4 presents a histogram of the similarity scores between Bitcoin and other cryptocurrencies. The five most similar coins to Bitcoin, based on return correlation in 2021, are Wrapped-Bitcoin (0.99), Huobi-Bitcoin (0.86), Litecoin (0.78), Ethereum (0.76), and Compound-Ether (0.74). As Wrapped Bitcoin replicates Bitcoin and Huobi-Bitcoin is backed by Bitcoin, their high return similarity is expected. Additionally, large market cap coins, such as Litecoin and Ethereum, also exhibit a strong return correlation with Bitcoin in 2021.

Likewise, the five cryptocurrencies most similar to Bitcoin based on whitepaper similarity are Nano (0.42), Ethereum (0.40), Zilliqa (0.37), NEM (0.37), and Tezos (0.36). We found that features such as "proof of work," "broadcast," "chain," and "block" contributed to the similarity scores with Bitcoin. Thus, whitepaper analysis captures the technological similarities between the cryptocurrencies.

Employing the fundamental similarity score, we identified the most similar coins to Bitcoin in 2021 as Wrapped-Bitcoin (0.99), Ethereum (0.99), Maker (0.96), Huobi-BTC (0.88), and Staked-Ether (0.87). Given that fundamental features are derived from returns and market capitalization, the scores closely resemble the return correlations.



(a) Return Correlation Score 2021

(b) Whitepaper Similarity

(c) Fundamental Similarity Nov 2021

Figure 3-4: Distribution of Bitcoin Similarity with other Currencies

Next, we conduct a quantitative test to assess the similarity scores. Let the feature vectors of coins i and j be represented by f_i and f_j, respectively. We perform the following regression:

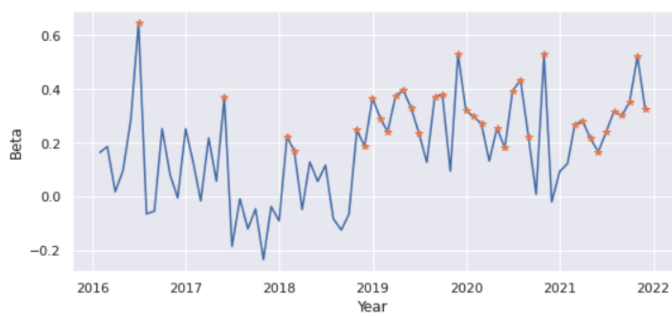
$$corr(r_{i,t+1:t+x}, r_{j,t+1:t+x}) = \alpha + \beta * similarity(f_{i,t-y:t}, f_{j,t-y:t}) \tag{3.1}$$

where $r_{i,t+1:t+x}$ is the returns time series of the coin i from time $t+1$ to $t+x$. $f_{i,t-y:t}$ represents the feature of coin i using the data from time $t-y$ to t. The above regression measures if the similarity measure can capture the future return correlation between the coins. A high, positive and significant β indicates that similarity measure is predictive of future similarity. The regression's universe comprises the top 100 coins by market capitalization.

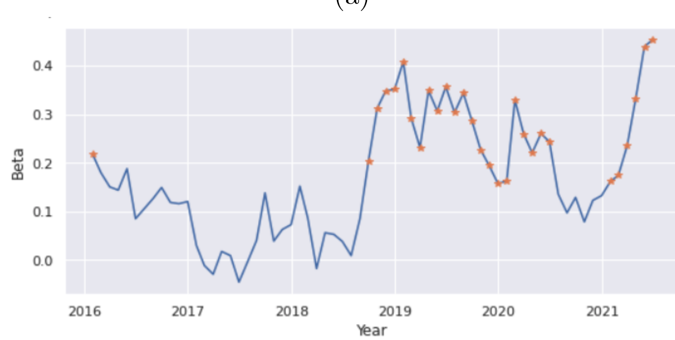
Whitepaper Similarity

Since whitepapers do not change over time, we can eliminate the subscript t from Equation 3.1 and run a linear regression independent of time. Our findings reveal that $\beta = 0.09(\pm 0.02)$, which implies that an increase in the similarity of coins by one unit will result in a 9% average increase in the correlation between a pair of coins.

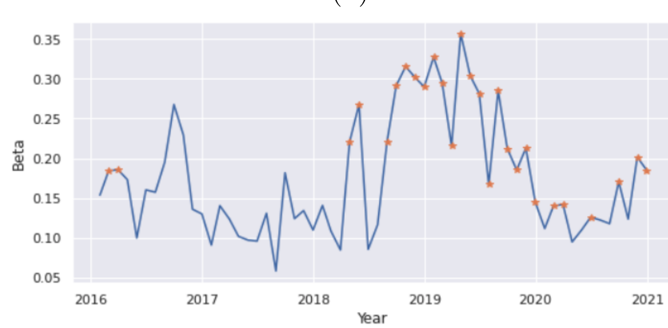
We also compute the time series β to examine the time-varying dynamics of β. In Equation 3.1, the forward return correlation can be calculated for different windows. In this analysis, we consider three windows: 1, 6, and 12 months. As a result, we assess the ability of whitepaper similarity to predict short-term return correlation (1 month) as well as long-term return correlation (12 months). Figure 3-5 displays the time-varying beta, with significant β values marked with an asterisk (*). We observe that whitepaper similarity is more predictive of return correlation in recent years (with a higher number of significant β values). This pattern holds true for different return correlation time horizons (1, 6, and 12 months), indicating that the whitepaper similarity's predictive power is consistent across these horizons. The time series average of β remains at 0.17 consistently across the return correlation prediction horizon.



(a)



(b)



(c)

Figure 3-5: Time Varying β for the Whitepaper Similarity Score.

Return Similarity

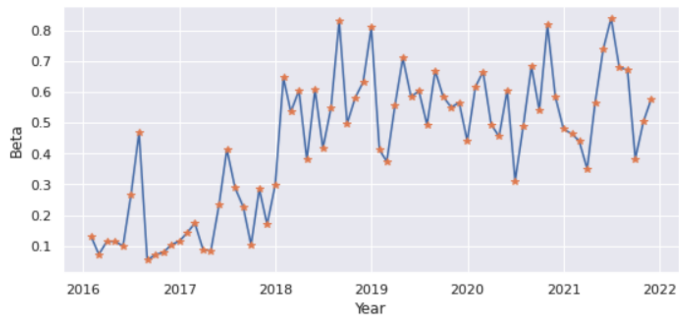
Another approach to computing similarity involves using the historical returns available for cryptocurrencies. By incorporating historical returns correlation as the similarity measure in Equation 3.1, we assess the predictive power of historical returns correlations in forecasting future return correlations. Historical returns similarity can be calculated using different time windows, such as 1 month (short-term similarity), 3 months (medium-term similarity), and 12 months (long-term similarity). These measures can then be employed to predict future returns across three horizons: 1 month (short-term similarity), 6 months (medium-term similarity), and 12 months (long-term similarity).

Figure 3-6 displays the time series β for the regression represented by Equation 3.1, with significant values marked by an asterisk (*). Our findings reveal that return similarity scores consistently predict the future resemblance of cryptocurrencies. The predictive power is lower in earlier years and higher in later years. The β values reach as high as 0.8, indicating that a one-unit increase in historical return similarity can lead to up to an 80% increase in future return correlation. The time series average of β ranges from 0.43 (for 1-month similarity) to 0.58 (for 3- and 12-month similarity).

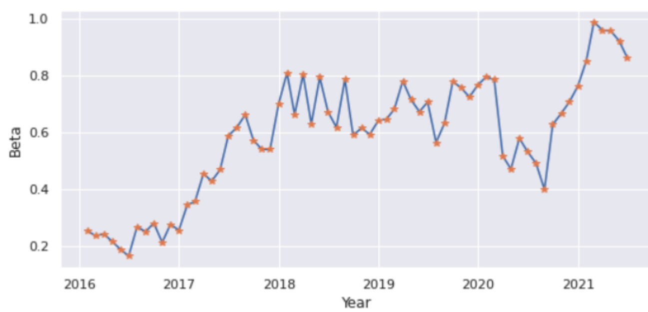
Fundamental Similarity

The third approach to measuring similarity between crypto-currencies involves using fundamental factors. The method for computing fundamental factors is described in Section 3.2.2. By employing the factor measures, we calculate similarity as the cosine similarity between the vectors of factor values for two coins and estimate the regression β outlined in Equation 3.1.

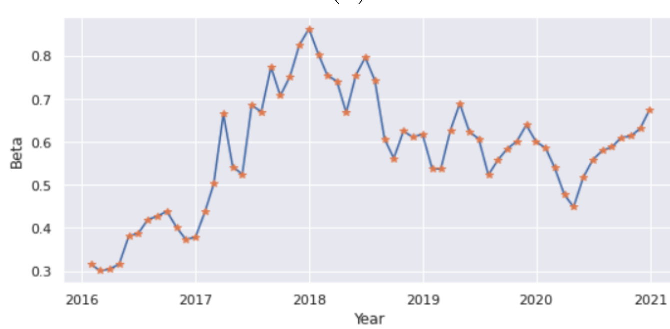
Figure 3-7 presents the time-varying β for regression (3.1). Similar to whitepaper and return similarity, the β is smaller in earlier years and larger in more recent years. The fundamental β remains significant throughout the historical period, but it is smaller than the return similarity β and larger than the whitepaper β. In summary, the fundamental similarity also captures the future correlation between



(a)



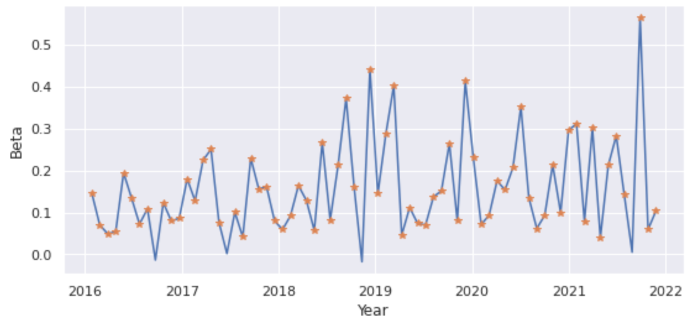
(b)



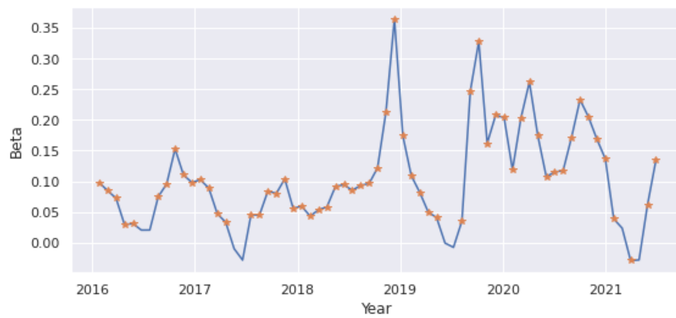
(c)

Figure 3-6: Time Varying β for the Return Similarity Score.

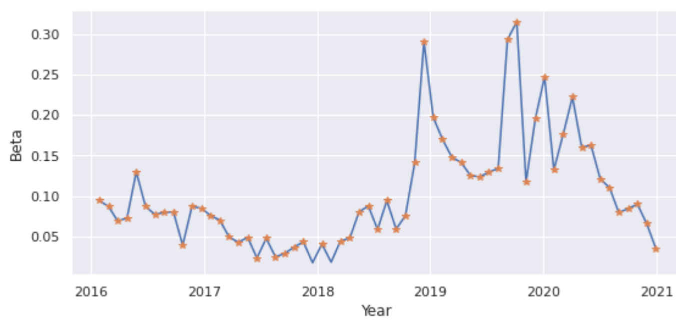
cryptocurrencies.



(a)



(b)



(c)

Figure 3-7: Time Varying β for the Fundamental Similarity Score.

Long Short Portfolios

Another test we conduct to assess the utility of the similarity score involves constructing long-short hedged portfolios. Given a cryptocurrency i and its peers j, k, l, m, n, we create a portfolio that takes a long position in coin i and equal-weighted short position in peers j, k, l, m, n. The returns of the hedged portfolio are given by:

$$r_{i_{L-S},t} = r_i - \frac{r_j + r_k + r_l + r_m + r_n}{5} \tag{3.2}$$

where, $r_{i_{L-S},t}$ represents the returns of the long-short hedged portfolio at time t. We calculate the volatility of the $L-S$ portfolio for every cryptocurrency and the volatility of individual coins, then measure the reduction in volatility. If the similarity scores effectively capture the co-movement of the coins, we should observe a statistically significant decrease in the volatility of single coins and the $L-S$ portfolio. The returns and fundamental similarity vary over time; therefore, we rebalance the portfolios at different time windows: 1 month, 6 months, and 12 months.

In Table 3.2, we present the average daily volatility of single-coin portfolios and Long-Short portfolios constructed using peers. The peers are obtained from the similarity score measure. We observe that Long-Short portfolios exhibit lower volatility[3]. This finding implies that similarity scores have utility in reducing the volatility of single-coin portfolios.

	1 month	6 month	12 month
Coins Vol. (Whitepaper)	6.75	7.02	7.09
L-S Vol. (Whitepaper)	5.79	5.98	6.11
Coins Vol. (Returns)	8.75	9.03	9.22
L-S Vol. (Returns)	8.22	8.37	8.52
Coins Vol. (Fundamental)	7.28	7.89	8.15
L-S Vol. (Fundamental)	6.31	6.81	7.01

Table 3.2: Volatility (monthly in percentage) of Coins and Long-Short Portfolio Constructed Using Peers

Clustering

Given the similarity scores for cryptocurrencies, we construct clusters of similar cryptocurrencies using hierarchical clustering algorithms. These clusters, which can be considered as currency groups or risk factors, group together similar cryptocurrencies.

[3]The decrease in volatility is statistically significant with p-val 0

To illustrate the clusters, we use daily return correlations for the year 2021 to construct clusters and present them in Figure 3-8. From the clusters, we can observe that Ethereum-based swap tokens like Uniswap and Sushi are grouped together. Similarly, exchange tokens like Binance Coin and KuCoin, as well as meme coins like Dogecoin and Shiba Inu, are grouped together. Therefore, the clusters of such coins may represent risk factors like exchange coins, meme coins, etc.

Considering each cluster as a risk factor, we compute the returns of each cluster factor. For the clustering, we have fixed the number of clusters to six. To illustrate the cluster factor properties, we compute the returns of the Bitcoin cluster (the cluster that contains Bitcoin) and calculate the statistical properties of the Bitcoin cluster. We have different versions of similarity scores: historical 1/6/12 month returns correlation, whitepaper similarity, and fundamental factor. For comparison purposes, we include the market index returns and Bitcoin returns.

Figure 3-9 presents the log cumulative returns of the Bitcoin cluster factor using factors obtained from different similarity score measures. The annualized mean returns and Sharpe ratio are presented in Table 3.3. We find that the Bitcoin cluster factors (except for the factor obtained from the fundamental similarity score) have higher annualized returns and Sharpe ratio compared to the market or Bitcoin.

Similarity Measure	Mean Returns	Sharpe Ratio
1 month return correlation	252	1.63
6 month return correlation	333	1.19
12 month return correlation	166	1.75
Whitepapers	188	1.64
Fundamental	120	1.33
Market Index	121	1.46
Bitcoin	110	1.45

Table 3.3: Mean Returns and Sharpe Ratio of Equal Weighted Cluster Factor Portfolios and Market Index.

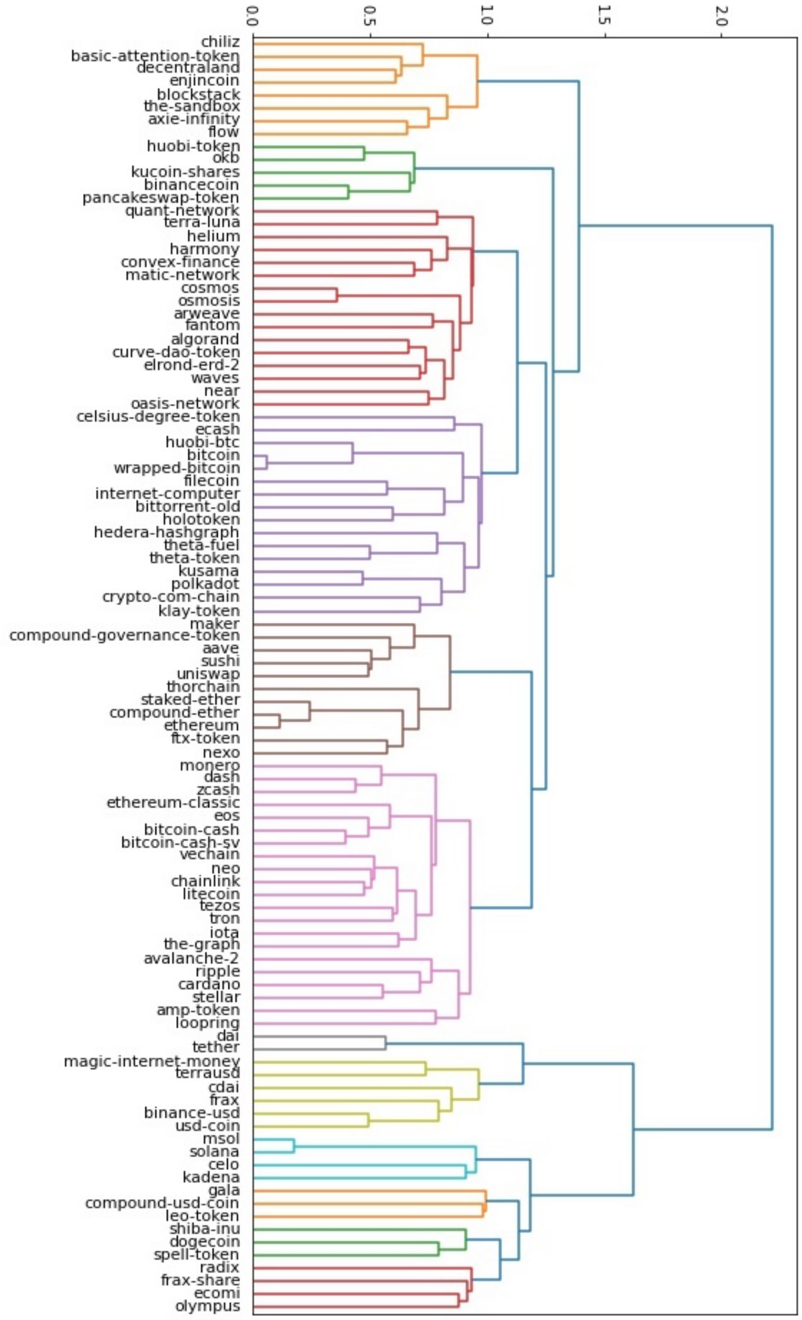


Figure 3-8: Clustering of Cryptocurrencies based on the Return Similarity for Year 2021.

Figure 3-9: Log Cumulative Returns of the Bitcoin Cluster Factor

3.4 Conclusion

In this study, we develop a similarity scoring system for cryptocurrencies and establish a peer grouping method that clusters similar currencies based on historical returns, whitepapers, and fundamental factors. Our results demonstrate that the similarity scores effectively capture the future return correlation of the coins. Furthermore, we show that these scores can be utilized to reduce the volatility of individual stock portfolios. The clusters generated using the similarity scores may serve as risk factors and find applications in risk management. However, one limitation of the peer grouping system is the difficulty in assigning human-interpretable summaries to the clusters.

Part II

Finance For Impact

Chapter 4

Finance For Impact: The Reaction of Sponsor Stock Prices to Clinical Trial Outcomes

In this chapter, we perform an event study analysis that quantifies the market reaction to clinical trial result announcements for 13,807 trials from 2000 to 2020, one of the largest event studies of clinical trials to date. We first determine the specific dates in the clinical trial process on which the greatest impact on the stock prices of their sponsor companies occur. We then analyze the relationship between the abnormal returns observed on these dates due to the clinical trial outcome and the properties of the trial, such as its phase, target accrual, design category, and disease and sponsor company type (biotechnology or pharmaceutical). We find that the classification of a company as "early biotechnology" or "big pharmaceutical" had the most impact on abnormal returns, followed by properties such as disease, outcome, the phase of the clinical trial, and target accrual. We also find that these properties and classifications by themselves were insufficient to explain the variation in excess returns observed due to clinical trial outcomes.

4.1 Introduction

Many factors are known to influence the stock prices of companies. They include earnings announcements[110], dividends[131], corporate actions[221], the release of new products [53], etc. For biopharmaceutical companies, apart from these general factors, clinical trial outcomes are critical since they play a major role in determining the future revenue, growth, and valuation of the company sponsoring the trial. For investors in this sector, clinical trial results are the predominant factor determining their return on investment. Therefore, quantifying the impact of clinical trial results on the stock prices of the companies is of great financial significance.

Apart from the goals of investors, however, the study of the movement of stock prices before the release of public information is also relevant for regulators, as an unexpected rise or fall of stock prices just *before* the release of news may be indirect evidence of insider trading[107, 224]. In addition, the study of excess returns after an event can be used to gauge the efficiency of markets. The Efficient Markets Hypothesis states that share prices fully reflect all available information about a company[102]; hence, upon the public release of new information, the price change should reflect this information. Therefore, the excess returns after an event can be used to gauge the degree of market efficiency and the speed with which the market incorporates new information.

On May 18, 2020, Moderna, Inc., announced its positive interim phase 1 trial data for its mRNA vaccine SARS-CoV-2, which led to its stock rising by 35% in a single day. Similarly, when Pfizer announced the interim results of its phase 3 SARS-CoV-2 vaccine, its stock rose by 9.2% on the announcement day. But the impact on stock prices from the phase 3 trial results by Pfizer was considerably smaller than the impact of the release of phase 1 trial data by Moderna. This observation raises multiple questions, including: what impact do clinical trial outcomes have on the stocks of their sponsoring companies? What properties of clinical trials influence these stock prices? And what kind of companies are more affected by the success or failure of a clinical trial? The answers to these questions will assist investors in their management of risk, and allow them to anticipate the impact of clinical trial results on their portfolio.

In this chapter, we answer these questions by performing an event study analysis for 13,807 clinical trial outcomes using data from 2000 to 2020 for 379 publicly traded U.S. companies. To the best of our knowledge, this is the largest event study analysis of clinical trial outcomes. We analyze the abnormal returns (defined as excess returns relative to a linear factor model described in Section **Methodology**.) for the multiple days around event (the day before the event, the day of event, the day after the event). We restrict our analysis to a short time window around the event to reduce the impact of confounding events from affecting our analysis.

To perform the study, we first identify the dates most relevant to clinical trial announcements on which the largest impact on the stock prices of their sponsor companies occur. Next, we observe the variation in abnormal returns across trials with different outcomes and different trials with same outcome, and investigate the factors responsible for this variation. We then develop a model relating the observed abnormal returns to the factors of a clinical trial, including trial phase, outcome type, target accrual, disease therapeutic ID, trial design, and sponsor company classification. We call this analysis a return attribution model, since it attributes the abnormal returns to various clinical trial factors. We perform this attribution for abnormal returns over the days immediately prior to and after the event. Our model can be used to measure the impact of the results of clinical trials on the returns of their sponsor company by using the multiple factors associated with a clinical trial.

Through this analysis, we obtain several insights. For individual clinical trials, we find that the sponsor company classification had the greatest impact on abnormal returns, with early biotechnology companies having greater movement in stock prices than large pharmaceutical companies when trial results were declared. We also find that factors like phase 2/3 and 3 trials, the disease categories of genitourinary and ophthalmology, and placebo-controlled trials had the largest abnormal returns in their respective categories, while phase 1 and 1/2 trials, the disease categories of autoimmune/inflammation and endocrinology, and multiple ascending dose studies have the minimum abnormal returns in their respective categories. Furthermore, we observe a positive correlation between the total target accrual of the trial and the stock's abnormal returns. Finally, our abnormal return attribution model finds that clinical trial properties and sponsor company classification were not sufficient to explain differences in the excess returns across different trial outcomes.

Our main contributions can be summarized as follows:

- With 13,807 clinical trial outcomes from 2000 to 2020, the scope of our event study analysis is orders of magnitude larger than earlier studies, yielding more accurate inferences regarding the impact of clinical trial outcomes on sponsor company stock prices.

- We measure the relationship between the properties of clinical trials (phase type, outcome, disease category, target accrual, trial design) and the classification of their sponsor companies as biotech or pharma on the abnormal returns of sponsor company stock, which has not been extensively explored in prior event studies of clinical trials.

- A robust statistical model was developed that estimates the average abnormal returns due to a clinical trial's outcome given the properties of the clinical trial and its sponsor company.

In the **Data** and **Methodology** sections, we describe the details of the dataset and the methodology of our event study, summarized more concisely in Figure 4-1. In the **Results and Discussion** section, we present the results and our findings from the study, while the **Robustness Checks** section provides a summary of the tests we perform to validate our findings. We conclude with a discussion about potential extensions.

4.2 Literature Review

The event study is a widely used method of analysis to quantify the financial effects of an event on the price movements of a stock. One of the earliest published event

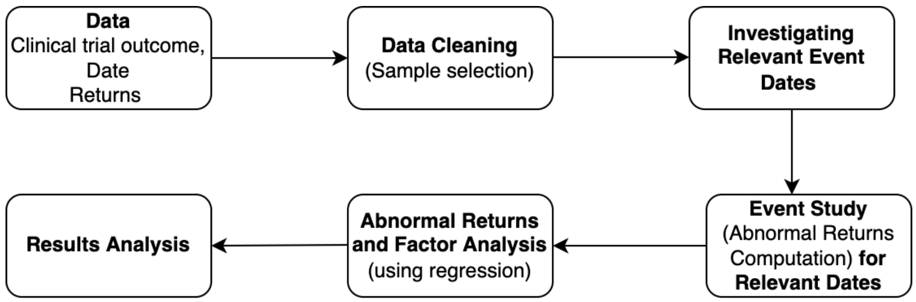


Figure 4-1: Steps involved in this event study, from the initial data collection to the analysis of our results.

studies was a 1933 student performed for measuring the impact of a stock split [89]. A study by *Fama et. al.*[103] proposed a methodology that is in essence the same as that used today, although over the years, it has been modified and refined for different use cases. For example, [42] deals with the issues in performing an event study for daily financial data. A number of survey papers [160, 198, 69] have summarized the econometrics of the event study and explored the evolution of its methodologies.

A variety of types of events have been analyzed in the literature, including mergers and acquisitions [208, 36, 55], earnings announcements ([141, 149], news headlines [52], financial reporting [92, 271], and stock messages on the Internet [12]. A detailed review of all the existing event studies in the literature is beyond the scope of this work. Instead, we focus on the event study literature that involves clinical trial results, FDA approval, drug failure, product development, and patent application.

An event study analysis of 24 clinical trial outcomes [146] found that there were asymmetric market reactions to positive and negative outcomes, i.e., underperformance due to negative events was greater in magnitude than the reaction to positive events. Unlike our findings, they conclude that the magnitude of abnormal returns did not differ by phase or firm classification. Another event study [224] consisting of 98 phase 3 trials and 49 products undergoing FDA advisory panel review sought evidence of insider trading, and found indirect evidence that it was common within the biotechnology sector. A study on the long-term returns around clinical trial results (that is, within 120 days of the event) focused on oncology drugs [243], while another [145] analyzed the market overreaction to clinical trial outcome. Focusing on biotechnology product development for 43 phase-3 trial outcomes, [267] analyzed the relationship between abnormal returns and company size, R&D expenditure, and type of result, while [90] analyzed the impact of drug features (e.g., disease and application type) and company features (e.g., size) on the value of abnormal returns.

Event study analyses have also been conducted on FDA approvals [266], pharmaceutical product development [254], drug failures [88], earnings surprises and product news [111], patent infringement [225], and patent application and publication dates in the pharmaceutical industry [263]. More recently, the influence of news of COVID-19 research and development on the biopharmaceutical industries has been analyzed [298], as has the impact of COVID-19 on the stock market[19] and the Indian pharmaceutical industry [23]. Figure C-2 in the Appendix summarizes these literatures.

Our work focuses specifically on clinical trial outcomes. While some earlier work focuses on clinical trial results, their scope is limited to due to their very small sample sizes (typically fewer than 150 events), while the events span a much shorter time period. Neither does this earlier work analyze the relationship of clinical trial features like phase, trial design, or accrual on abnormal returns. With over 13,000 clinical trial outcomes spanning a 21-year period, the scope of our event study analysis is orders of magnitude larger than those of previous studies, which translates into more accurate inferences about the impact of clinical trial outcomes on stock's price movements. Moreover, we also analyze how the impact of clinical trial outcomes on sponsor company stock price movements differs over time.

4.3 Data

We use Citeline data provided by Informa Pharma Intelligence for details of the clinical trials, their outcomes, and their properties. The dataset combines individual clinical trial information from *Trialtrove* and drug approval data from *Pharmaprojects*. It is one of the world's most comprehensive dataset of global clinical trials and drug development pipelines. The dataset consists of trials sponsored by government, academia, cooperative groups, and industry. We remove trials that are not sponsored by industry. Next, we filter out clinical trials based on multiple criteria: missing event dates, missing reasons for termination, non-availability of returns data for sponsor companies, and ambiguity about the sponsors. Ambiguity in sponsorship refers to unclear information about the sponsor company in the dataset. Furthermore, we filter trials based on valid outcomes or early termination reasons, including only those trials that belong to one of the following categories: safety/adverse effect, lack of efficacy, early positive outcome, positive outcome/primary endpoints met, and negative outcome/primary endpoints not met. To analyze the different outcome types more intuitively, we group trials from different categories as follows: a negative finding includes primary endpoints not met and lack of efficacy; a positive finding includes an early positive outcome and primary endpoints met; closed early includes a lack of efficacy and safety/adverse effect; and full maturity includes primary endpoints met, early positive outcomes and primary endpoints not met.

After all our filters, the final sample sizes of trials are presented in Table C.1 of the Appendix. We also include the trial distribution across different phases (1, 2, 3, 4, 1/2, 2/3, 3/4), sponsor company types (biotechnology vs pharmaceutical classification), therapeutic categories (9 categories), and trial design (28 in total). The filtering process resulted in a sample of 13,807 trials, and this process is described in more detail in Appendix C.1 and Figure C-1.

We obtain the sponsor company classification from [273], who classify companies using a k-nearest-neighbors clustering algorithm into the following categories on the basis of research and size of the company. Big Pharmaceutical (BP) company research traditionally includes (but is not limited to) small molecules, and BP companies have numerous products that reach patients. Small Pharmaceutical (SP) companies have fewer assets than big pharmaceutical companies, and tend to produce drugs for niche diseases. These companies also have products on the market. Early Biotechnology (EB) companies differ from pharma companies in that their production process typically uses living material, but they may also produce small molecules. EB companies may or may not have drugs in the market. Late Biotechnology (LB) companies are larger in size than EB companies, with a product in market. In the following work, we use the categories EB, LB, SP and BP to classify the companies.

In addition to Citeline dataset, we obtained daily stock returns of the trial sponsor companies from the Center for Research in Security Prices (CRSP). We merged the Citeline and CRSP dataset using fuzzy string matching and manual verification of the correct match on sponsor and company names.

4.4 Methodology

An event study formally consists of multiple components: event definition, stock selection, model definition, model estimation, empirical results, and interpretation. In our analysis, we first define the day of the 'event' (the trial outcome declaration date) from the relevant clinical trial dates. Following this step, the stocks of the sponsor companies are selected for the study. For this analysis, we use the Fama–French five-factor model to estimate abnormal returns (The analysis using other factor models is available in the Appendix) Finally, after calculating the abnormal returns, we attribute them to different properties of the clinical trials and sponsor companies.

Event Dates

For a given clinical trial, there are multiple important dates related to trial outcomes. They include:

- **Primary Completion Date**: The date when the final subject was examined or received intervention for data collection for evaluating the primary endpoints of a clinical trial.

- **Primary Endpoint Reported Date**: The earliest date of public report of results that address the primary endpoints of the trial. Reports of interim or preliminary results do not constitute a primary endpoint reported date.

- **Publish Date**: The earliest date of result publication, which includes interim and preliminary results.

- **Enrollment Close Date**: The date on which the study completed patient enrollment.

The most movement in the company returns should happen on the day when information about trial outcomes is released to the public for the first time. Later in our results, we will show how different dates are relevant for different trial outcomes.

Sample Selection

The companies sponsoring the trials are most directly affected by the outcomes of the clinical trials; hence, we select these sponsor companies for our analysis. In some cases, there is an ambiguity in sponsorship, which happens when sponsor companies merge with, or are acquired by, other companies during the course of trial. In such cases, we identify the correct sponsor company to use for the event study by looking at the merger or acquisition date.

Abnormal Returns Model

In most event studies, it is conventionally assumed that asset returns are jointly multivariate normal in distribution, and independently and identically distributed over time. These assumptions allow us to develop various simple models for the estimation of abnormal returns prior to, on, or after the day of the event. We use eventstudy python package for the estimation procedures, which are described in more detail in [47].

The Fama–French five-factor model [105] determines the relationship between different factor returns and security returns:

$$R_{i,t} = \alpha_i + \beta_{i,1} * R_{m,t} + \beta_{i,2} * R_{SMB,t} + \beta_{i,3} * R_{HML,t} + \beta_{i,4} * R_{RMW,t} + \beta_{i,5} * R_{CMA,t} + \epsilon_{i,t},$$
$$E[\epsilon_{i,t}] = 0, Var[\epsilon_{i,t}] = \sigma^2_{\epsilon_i}$$

where $R_{i,t}$ represents the returns of asset i at time t; $R_{m,t}$, $R_{SMB,t}$, $R_{HML,t}$,$R_{RMW,t}$, and $R_{CMA,t}$ are the factor returns at time t; and α_i, $\beta_{i,1}$, $\beta_{i,2}$, $\beta_{i,3}$, $\beta_{i,4}$, $\beta_{i,5}$, and $\sigma^2_{\epsilon_i}$

are the factor loadings or parameters.

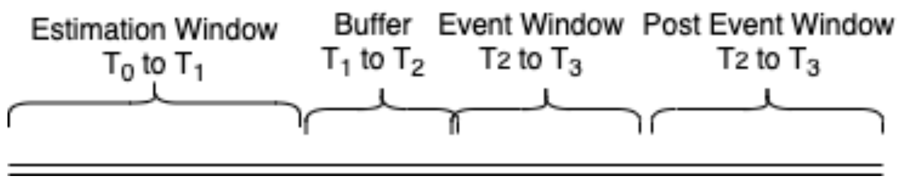


Figure 4-2: Timeline of the event.
T_0 to T_1 is the window used for model estimation. A buffer window is kept between the estimation and the event window to prevent any information leakage. The event window is T_2 to T_3, so that the day of the event $t = 0$ lies within the event window.

The timeline of an event is shown in Figure 4-2. The model parameters can be estimated using the observed returns of the security and factors during the estimation period. The estimation window in our study was 252 trading days (i.e., one year). Once the model's parameters are estimated, they are used to obtain the returns in the event window, $\widehat{R_{i,t}}$, and abnormal returns (AR) are calculated as the difference between the observed and estimated returns, $(R_{i,t} - \widehat{R_{i,t}})$. When multiple ($N$) events are analyzed together, the average abnormal returns (AAR) (we use the terms 'abnormal returns' and 'average abnormal returns' interchangeably in the remaining analysis.) are given by:

$$AAR = \frac{\Sigma AR}{N}.$$

Return Attribution

To measure the impact of clinical trial and sponsor company properties on the stock prices of the sponsor companies, we perform a regression of abnormal returns on these properties:

$$Y_i = \alpha + \beta * phase + \gamma * company_{type} + \delta * outcome + \eta * disease + \theta * targetAcc. + \zeta * design, \tag{4.1}$$

where the regressors are coded using one-hot encoding, except for $targetAcc.$, which is a continuous variable. The sign of abnormal returns of unsuccessful outcomes (e.g., a primary endpoint not met, an adverse effect, or a lack of efficacy) are inverted to make them comparable with positive outcomes. The regressors are:

phase:	clinical trial phase
$company_{type}$:	sponsor company type
outcome:	clinical trial outcome
disease:	therapeutic care classification associated with a trial
targetAcc.:	target accrual for a trial
design:	clinical trial design

Y_i represents the abnormal returns, and i indicates the various time windows around the event date for which the abnormal returns were calculated: $i = -1$ for the day prior to the event window; $i = 0$ for the day of the event; $i = 0$ to 1 for the day of and day after the event; and $i = 2$ for two days after the event.

Since Y_i are abnormal returns, which are estimated with some error, we use bootstrapping to compute the confidence intervals associated with the coefficients of the regressions.

4.5 Results and Discussion

In this section, we first present the identification of the most significant date for various clinical outcomes, then analyze the abnormal returns for a two-week period before and after the events, and discuss the relationship between abnormal returns and clinical trial properties. Finally, we include examples of the average abnormal returns due to clinical trial outcomes conditional on multiple factors.

Clinical Trial Outcome Dates

Depending on the outcome of the trial, different dates may be significant for clinical trials. We first determine the dates associated with clinical trials that have the largest impact on the stock price of the companies. To do so, we segregate trials on the basis of reason for termination and analyze them separately. We run a multi-event study for the trials with the following event dates: the primary completion date, the date when the primary endpoints were reported, the earliest date of result publication, and the enrollment close date. We present the abnormal returns on the day of the event, day 0 and the cumulative abnormal returns on day 0 and 1, along with the standard errors in Table 4.1.

Outcome	Day	Primary Completion	Endpoint	Publish	Close
Safety/Adverse Effect	0	**−0.75** (0.09)	0.08 (0.11)	−0.39 (0.09)	−0.40 (0.10)
	0,1	**−0.82** (0.12)	0.25 (0.15)	−0.72 (0.13)	−0.26 (0.14)
Lack of Efficacy	0	**−1.70** (0.09)	−0.28 (0.09)	−0.45 (0.07)	−1.58 (0.10)
	0,1	**−2.43** (0.13)	−0.32 (0.13)	−0.47 (0.11)	−2.37 (0.15)
Early Positive Outcome	0	**5.31** (0.20)	0.03 (0.24)	0.38 (0.25)	0.05 (0.21)
	0,1	**6.35** (0.28)	0.48 (0.33)	1.07 (0.35)	0.41 (0.30)
Primary Endpoints Met	0	−0.02 (0.02)	0.15 (0.02)	**0.46** (0.02)	0.04 (0.02)
	0,1	−0.01 (0.03)	0.12 (0.03)	**0.54** (0.03)	0.08 (0.03)
Primary Endpoints Not Met	0	−0.20 (0.04)	−0.33 (0.05)	**−1.97** (0.04)	0.06 (0.04)
	0,1	−0.19 (0.06)	−0.49 (0.07)	**−2.71** (0.06)	0.17 (0.06)

Table 4.1: Average abnormal returns for different outcome types on various significant dates

Average abnormal returns on day of the event (day 0) and cumulative average abnormal returns (day 0,1) for trials with different outcomes on significant dates with their standard error (bracket). Day 0 abnormal returns are in the first row and day 0,1 abnormal returns are in the second row for each outcome. We find that for trial outcomes of safety/adverse effect, lack of efficacy, and early positive outcome, the highest abnormal returns are obtained on the primary completion date. For completed trials with outcomes of primary endpoint met or not met, the highest magnitude of abnormal returns are obtained on the earliest date of result publication. The maximum absolute abnormal returns are in bold.

We observe that the magnitude of average abnormal returns for an early positive outcome are greater on the primary completion date than for any other date. This implies that, as soon as the data from the final subject has been collected and the results appear promising, companies will release the results, so that the largest abnormal returns are observed on the primary completion date. However, when the outcome is that the primary endpoint is met or not met, we find that the value of average abnormal returns is greatest on the earliest date of result publication (the publish date). On this date, companies release their interim results after a thorough study of the data collected thus far.

Trials that show a safety/adverse effect or lack of efficacy have the largest magnitude of abnormal returns on their primary completion date, but these are comparable to the abnormal returns obtained on the enrollment close date and the publish date. Trials that show a safety/adverse effect or lack of efficacy outcome are terminated early, and development is stopped. In the case of a safety/adverse effect, for a majority of trials, the primary completion date is earlier than the publish date. Since details about the termination of the trial are available on the primary completion date, we therefore observe the largest abnormal returns on the primary completion date. For trials that show a lack of efficacy, we observe the same primary completion and enrollment close date for many trials. This explains the higher abnormal returns on the enrollment close date.

For the remainder of this study, we use the primary completion date as the event date for the following outcomes: a safety/adverse effect, a lack of efficacy, and an early positive outcome. These outcomes constitute 9% of the dataset. For the remaining results, when primary endpoints are met or not met, we use the publish date as the event date. These outcomes constitute 91% of the dataset.

Abnormal Returns

In Figures 4-3, 4-4, and 4-5, we include the average abnormal returns and the cumulative average abnormal returns for trial outcomes for the two-week trading period before and after the event date. As expected, we observe high average abnormal returns on the day of the event (day 0 and day 1). For trials that declare their results after trading hours, the movement in stock prices is expected to happen on the day after the result declaration date. For unsuccessful trial outcomes such as a safety/adverse effect, a lack of efficacy, or primary endpoints not met, we observe cumulative abnormal returns decreasing during the two-week period after the event date.

In Figures 4-3, 4-4, and 4-5, we observe variation in average abnormal returns. Even within each outcome type, we find a huge spread in the abnormal returns of

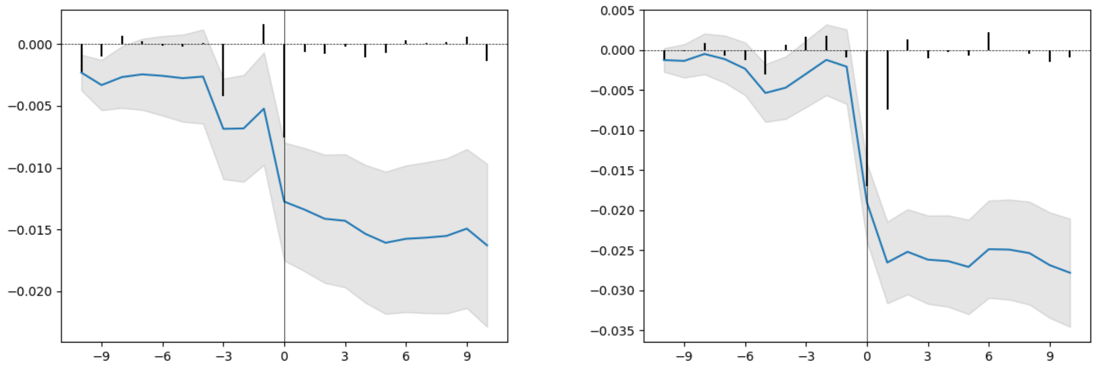


Figure 4-3: Left: Safety and Adverse Effect , Right: Lack of Efficacy

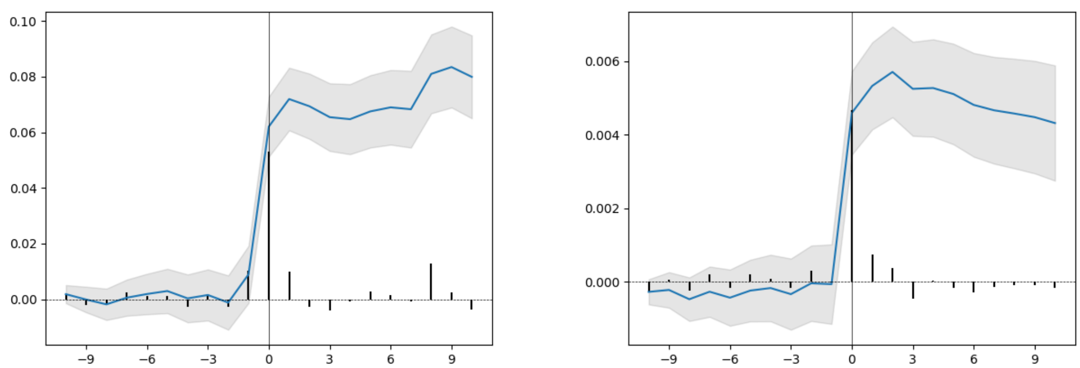


Figure 4-4: Left: Early Positive Outcome, Right: Primary Endpoints Met

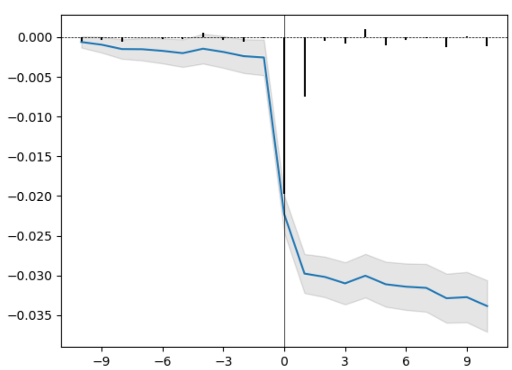


Figure 4-5: Primary Endpoints not met

Average abnormal returns and cumulative average abnormal returns with a 95% confidence interval for a multi-event study of trials with different outcomes. The x-axis represents the day (0 is the day of the event) and the y-axis represents the abnormal return values (scale: fraction).

individual events. For a safety/adverse effect, lack of efficacy and primary endpoints not met outcomes, the minimum value of abnormal returns for individual events is -80%, -90%, and -100% respectively. However, for the early positive outcome and primary endpoints met, the maximum value of abnormal return is 137% and 485%, respectively, which is quite different from the average abnormal returns. Such variation in abnormal returns for individual events within each outcome category may potentially be due to the properties of the clinical trials and sponsor companies involved in the trials, including the trial phase, company classification, type of result, therapeutic area of the disease, and trial design.

Abnormal Returns and Trial Properties

We conjecture that the impact of a trial outcome on the sponsor's stock prices is dependent on specific properties of a clinical trial and its sponsor company. From Table C.1, we can observe that the distributions of trials across different phases and sponsor company types are different. Hence, to measure the impact of these trial and company properties on the stock prices of the sponsor companies, we perform a regression of abnormal returns on these different properties, as given by Eq. (4.1). For trials with negative outcomes, we reverse the sign of the abnormal returns prior to regression.

In Tables 4.2–4.8, we present the abnormal returns for various properties (given by the regression coefficients) with their standard errors on days around the event, and analyze them in the following section. In Figure 4-6, we present the contribution of clinical trial properties on the average abnormal return of a company on the day of the event (day 0-1) in a bar plot. We then discuss each trial or company property and its abnormal return individually.

Clinical Trial Phase

Clinical trials are conducted in phases, with each phase designed to answer specific questions. Phase 1 is designed to evaluate the safety of the treatment, phase 2 is designed to evaluate the effectiveness of treatment in a small and carefully curated population, and phase 3 trials are generally designed to compare the effectiveness and safety in a larger and more heterogenous population against the existing standard treatment. Typically, therapeutics obtain FDA approval after a successful phase 3 trial, although they may go to a phase 4 trial following approval, in which the long-term side effects of the new therapeutic are studied. In many cases, different trial phases are combined to speed up the development and testing process.

Statistics for the abnormal returns for each phase category (keeping other factors constant) are given in Table 4.5. We observe that phase 2/3 trials have the greatest impact on the stock returns of sponsor companies, followed by phase 3 trials, while phase 1 and 1/2 trials have the least impact on abnormal returns. On the day of the event, day 0 and day 0–1, phase 2/3 trials have higher abnormal returns than phase 1/2 trials by 2.27% and 3.06%, respectively.

Phase 3 trials are generally costly and require significant capital[237], and the results of phase 3 trials determine whether a therapeutic can be commercialized and used by patients. This explains the higher abnormal returns for phase 3 and 2/3 trials. Even though phase 4 trials happen after FDA approval, negative findings from such trials raise questions about the long-term effectiveness of the therapeutic; hence, phase 4 trials had higher abnormal returns as well.

Company Type

A large number of biotechnology and pharmaceutical companies are involved in the development of new therapeutics as clinical trial sponsors. The success or failure of these clinical trials has an impact on the business of the sponsor company, leading to changes in its stock prices.

We present the statistics of abnormal returns by sponsor type in Table C.12. As observed, the abnormal returns for early biotechnology companies are the largest, followed by small pharmaceutical companies, with large pharmaceutical companies having the smallest abnormal returns. Early biotech companies had 8.32% and 10.95% higher abnormal returns than large pharmaceutical companies on the day of the event—day 0 and day 0–1—respectively.

Early-stage biotechnology companies tend to be smaller in size, and are involved in earlier stage discovery and development efforts for specific potential therapeutics. In general, they do not have FDA-approved products. The trial results for a sponsored therapeutic candidate thus have a large impact on the growth of the company. Hence, we observe higher values of abnormal returns for these companies. At the same time, large pharmaceutical companies have a higher market capitalization and tend to have multiple products on the market; hence, the results of clinical trial outcomes have the least impact on this category of companies. Similarly, late-stage biotechnology companies are established companies with FDA-approved products already on the market, and thus clinical trial results have less of an impact on their stock market returns.

Moreover, we observe that the sponsor company classification has the biggest impact on abnormal returns compared to other properties of clinical trials.

Outcome Type

As mentioned in earlier sections, the trials analyzed in our sample have very different outcomes. Table 4.1 shows the average abnormal returns for these different outcomes on various dates. As observed, the abnormal returns for unsuccessful outcomes are negative, while those for successful trials are positive. In the regression in Eq. (4.1), we invert the sign of unsuccessful outcomes to compare them directly to successful outcomes. Here, we analyze the magnitude of abnormal returns controlling for trial properties.

In Table 4.2, we present the abnormal returns by outcome type. We find that trial results with early positive outcomes have the largest abnormal returns, followed by unsuccessful outcomes like unmet primary endpoints and a lack of efficacy. The outcome of primary endpoints met have the lowest value of abnormal returns. The difference between abnormal returns for trials with an early positive outcome and trials with successful endpoints are 2.95% and 3.39%, respectively, on day 0 and day 0–1. The high abnormal returns for early positive outcomes may be due to the surprise factor associated with an earlier than anticipated declaration of results. Higher abnormal returns for trials with unsuccessful outcomes than for trials with successful outcomes (asymmetric market reaction) is consistent with earlier findings in the literature [146].

To analyze the different outcome types more intuitively, we create new groups from the dataset's original categories as follows: a negative finding includes primary endpoints not met and a lack of efficacy; a positive finding includes an early positive outcome and primary endpoints met; a finding of closed early combines a lack of efficacy and safety/adverse effect; and full maturity combines primary endpoints met, early positive outcomes and primary endpoints not met. We find that the magnitude of abnormal returns of clinical trials due to a negative finding is higher than a positive finding, consistent with earlier results in the literature [146]). Similarly, trials that are closed early have a higher magnitude of abnormal returns than trials reaching full maturity.

Target Accrual

Target accrual is the number of subjects that are planned to participate in the study. A larger target accrual signifies a larger investment by the company for the clinical trial, and a treatment tested on a larger number of subjects has more reliable results. We include the coefficient for target accrual in Table 4.3. We observe that trials with larger targeted accruals have a higher value of abnormal returns. More specifically, an increase in the targeted accrual of a trial by 1,000 subjects, keeping the other properties constant, produces a higher value of abnormal returns by 0.158% on the day of the event.

Disease

Clinical trials test potential treatments for diseases belonging to many different therapeutic areas. In our data, these include: genitourinary, ophthalmology, vaccines (for infectious disease), CNS, oncology, cardiovascular, infectious disease, autoimmune/inflammation, and metabolic/endocrinology. By including the disease type as one of the regressors, we are able to analyze the impact of disease category on a company's abnormal stock returns.

In Table 4.7, we include the average abnormal returns by disease category. We find that genitourinary diseases have the largest abnormal returns, (the higher abnormal returns are due to multiple trial outcomes following the investigation of BioSante Pharmaceuticals), followed by ophthalmology and vaccine trials. The abnormal returns are smallest for endocrinology and autoimmune disease trials. It is widely observed that oncology drugs are best-sellers; however, contrary to this stylized fact, we do not observe a high level of abnormal returns for oncology trial outcomes, despite oncology trials being the largest disease category in our dataset. Nevertheless, the abnormal returns for oncology trials are higher than those for infectious disease, autoimmune, and endocrinology trials.

Design

There are multiple ways of designing clinical trials, controlling for cost, to establish the effect of a potential new therapeutic [100]. This trial design depends on the characteristics of treatment, clinical endpoints, trial phase, the availability of a control group, funding, and other properties.

In Table 4.7, we present the abnormal returns for different trial designs (in our regression, we exclude design categories that have fewer than 100 trials). We find that placebo-controlled trials have the largest abnormal returns, followed by pharmacokinetic, fixed dose, adaptive, and safety trials. Meanwhile, designs like single and multiple ascending dose, pharmacodynamic, and non-inferiority trial designs have the smallest abnormal returns. A placebo-controlled, double-blind trial design is considered the gold standard in determining the effect of a treatment and used for designing phase 3 trials. This potentially explains the higher abnormal returns of placebo-controlled trials.

We also observe that competing trial designs have varying abnormal returns. For example, for an open-label trial design, where the subject and researchers know that the subject is receiving a treatment, compared to a double-blinded trial design, where neither the subject nor the researchers know who is receiving a treatment or a placebo, the abnormal returns are not statistically different. Meanwhile, the single ascending dose trial design has higher abnormal returns than the multiple ascending dose design. Similarly, the superiority trial design (showing that a new method is better than an active control or placebo) has higher abnormal returns than non-inferiority trial designs (showing that a new method is not inferior to an active control or placebo).

Stock Movement Given Trial Properties

As discussed earlier, the linear factor model for abnormal stock returns in Eq: 4.1 can be used by investment managers to obtain average abnormal returns for the stock of a sponsor company at the declaration dates of a clinical trial result. This is demonstrated by the following examples.

The regression coefficients from Tables 4.2–4.8 can be used to compare the abnormal returns of clinical trials with different properties. For example, phase 2/3 trials sponsored by early-stage biotech (EB) companies with an early positive outcome have 13.54% higher absolute abnormal returns than phase 1/2 trials with a big pharmaceutical (BP) company sponsor and an outcome of primary endpoints met on the day of the event. Similarly, the cumulative abnormal returns for two days (day 0–1) for a clinical trial with an early positive outcome (2.56%) and a target accrual of one thousand subjects (0.134%), sponsored by an early biotechnology company (6.33%), at phase 3 of the trial (1.60%) for a treatment for genitourinary disease (2.85%) using a double-blinded, placebo-controlled trial design (1.23%–0.33%) are 18.49%. This information has several potentially useful applications for investment managers and other stakeholders.

The unexplained average abnormal returns after the regression are presented in Table 4.8. As expected, the average of the unexplained abnormal returns are not statistically significant prior to the day of the event (day -1) or after the event (day 2). The value of α increases from 2.83% on the event date (day 0) to 4.10% for the day after the event (including the event date). The higher value of unexplained abnormal returns (α) implies there are factors apart from the ones included in this study that might explain the abnormal returns of these companies.

We find that abnormal returns on the day before the event ($\delta_{-1}, \theta_{-1}, \gamma_{-1}, \beta_{-1}, \eta_{-1}, \zeta_{-1}, \alpha_{-1}$) are not significant. Similarly, the abnormal returns are not significant on the day after the event (δ_2, $\theta_2, \gamma_2, \beta_2, \eta_2, \zeta_2, \alpha_2$), implying that markets incorporate the information related to clinical trials, and the sponsor stock prices stabilize.

We also analyze the pattern of abnormal returns of the clinical trial outcomes over time. The discussion and results are included in the Appendix (in the section *Abnormal Returns Over Time* and Table C.2).

Table 4.2: Outcome types

Outcome	δ_{-1} (SE)	δ_0 (SE)	δ_{0-1} (SE)	δ_2 (SE)
Early Positive Outcome	0.18 (0.17)	2.21 (0.17)	2.40 (0.24)	0.08 (0.17)
Lack of Efficacy	0.09 (0.09)	0.57 (0.09)	1.31 (0.12)	-0.23 (0.09)
Primary Endpoints Not Met	-0.01 (0.06)	0.79 (0.06)	1.30 (0.09)	0.1 (0.06)
Safety/Adverse Effect	-0.22 (0.09)	-0.01 (0.09)	0.07 (0.13)	0.11 (0.09)
Primary Endpoints Met	0.0 (0.05)	-0.74 (0.05)	-0.99 (0.08)	0.0 (0.05)
Lack of Efficacy	0.09 (0.08)	1.17 (0.08)	1.99 (0.11)	-0.17 (0.08)
Negative Finding	-0.05 (0.05)	1.27 (0.05)	1.81 (0.07)	0.15 (0.05)
Positive Finding	0.00 (0.05)	-0.08 (0.05)	-0.24 (0.06)	0.05 (0.05)
Closed Early	0.00 (0.06)	1.33 (0.07)	2.11 (0.09)	-0.05 (0.06)
Full Maturity	0.02 (0.05)	0.64 (0.05)	0.85 (0.08)	0.08 (0.05)

Table 4.3: Target Accrual

	θ_{-1} (SE)	θ_0 (SE)	θ_{0-1} (SE)	θ_2 (SE)
Target Accrual	0.016 (0.015)	0.158 (0.015)	0.144 (0.022)	0.021 (0.015)

Table 4.4: Company sponsor category

Company Type	γ_{-1} (SE)	γ_0 (SE)	γ_{0-1} (SE)	γ_2
Early-stage Biotech	−0.33 (0.20)	4.65 (0.2)	6.31 (0.28)	0.36 (0.20)
Small Pharma	0.12 (0.15)	1.63 (0.15)	1.69 (0.22)	0.12 (0.15)
Late-stage Biotech	0.32 (0.15)	0.31 (0.15)	0.73 (0.21)	−0.21 (0.15)
Big Pharma	−0.02 (0.09)	−3.67 (0.09)	−4.64 (0.12)	−0.08 (0.09)

Table 4.5: Phase categories.

Phase	β_{-1} (SE)	β_0 (SE)	β_{0-1} (SE)	β_2 (SE)
2/3	−0.15 (0.11)	1.63 (0.11)	1.95 (0.16)	0.09 (0.11)
3	0.03 (0.06)	0.99 (0.05)	1.59 (0.08)	−0.04 (0.06)
4	−0.01 (0.06)	0.39 (0.06)	0.73 (0.08)	0.04 (0.06)
3/4	−0.13 (0.15)	0.32 (0.16)	0.68 (0.22)	−0.10 (0.16)
2	0.05 (0.06)	0.30 (0.06)	0.47 (0.08)	0.14 (0.06)
1	0.15 (0.08)	−0.33 (0.08)	−0.28 (0.12)	0.06 (0.08)
1/2	−0.07 (0.10)	−0.64 (0.10)	−1.11 (0.14)	0.07 (0.10)

Table 4.6: Disease therapeutic category

Disease	η_{-1} (SE)	η_0 (SE)	η_{0-1} (SE)	η_2 (SE)
Genitourinary	0.09 (0.22)	1.51 (0.22)	3.13 (0.32)	0.25 (0.22)
Ophthalmology	0.10 (0.16)	0.33 (0.16)	0.81 (0.23)	0.11 (0.16)
Vaccines (Infectious Disease)	0.04 (0.12)	−0.09 (0.12)	0.35 (0.16)	0.09 (0.12)
CNS	0.07 (0.11)	0.25 (0.11)	−0.18 (0.16)	0.04 (0.11)
Oncology	−0.02 (0.11)	−0.14 (0.11)	−0.43 (0.15)	−0.10 (0.11)
Cardiovascular	0.09 (0.10)	−0.01 (0.10)	−0.65 (0.15)	0.04 (0.10)
Infectious Disease	−0.02 (0.09)	−0.49 (0.09)	−0.76 (0.13)	−0.13 (0.09)
Autoimmune/Inflammation	0.07 (0.11)	−0.39 (0.11)	−0.81 (0.15)	−0.04 (0.11)
Metabolic/Endocrinology	0.02 (0.10)	−0.51 (0.10)	−1.00 (0.14)	0.03 (0.10)

Table 4.7: Clinical Trial Design

Design	ζ_{-1} (SE)	ζ_0 (SE)	ζ_{0-1} (SE)	ζ_2 (SE)
Placebo Control	-0.09 (0.07)	1.10 (0.07)	1.30 (0.10)	0.14 (0.07)
Pharmacokinetics	-0.03 (0.06)	0.46 (0.06)	0.99 (0.08)	-0.07 (0.06)
Fixed Dose	0.01 (0.14)	0.68 (0.13)	0.82 (0.19)	0.10 (0.13)
Adaptive	0.45 (0.24)	0.35 (0.23)	0.69 (0.33)	-0.32 (0.23)
Safety	-0.03 (0.06)	0.57 (0.06)	0.63 (0.08)	0.20 (0.05)
Non Interventional	0.02 (0.11)	0.44 (0.11)	0.40 (0.15)	-0.32 (0.11)
Active Comparator	0.06 (0.05)	0.05 (0.05)	0.30 (0.08)	0.08 (0.05)
Multiple Arm	0.05 (0.08)	0.25 (0.08)	0.26 (0.11)	-0.03 (0.08)
Observational	0.01 (0.09)	0.22 (0.09)	0.17 (0.13)	0.02 (0.09)
Single Arm	-0.03 (0.09)	0.14 (0.09)	0.00 (0.12)	-0.03 (0.09)
Randomized	-0.1 (0.08)	0.14 (0.08)	-0.04 (0.12)	-0.08 (0.08)
Efficacy	-0.04 (0.06)	-0.18 (0.06)	-0.06 (0.09)	-0.09 (0.06)
Immunogenicity	0.01 (0.08)	0.21 (0.08)	-0.26 (0.11)	0.02 (0.08)
Dose Response	-0.01 (0.09)	-0.07 (0.09)	-0.26 (0.12)	-0.03 (0.08)
Superiority	0.02 (0.10)	-0.11 (0.10)	-0.27 (0.14)	0.08 (0.1)
Cross Over	-0.01 (0.09)	-0.15 (0.09)	-0.30 (0.13)	-0.04 (0.09)
Double blind/blinded	0.06 (0.09)	-0.04 (0.09)	-0.33 (0.13)	-0.17 (0.09)
Open Label	0.03 (0.08)	-0.02 (0.08)	-0.34 (0.11)	-0.03 (0.08)
Pharmacodynamics	-0.01 (0.06)	-0.63 (0.06)	-0.85 (0.09)	0.03 (0.06)
Non Inferiority	-0.39 (0.08)	-0.64 (0.08)	-0.86 (0.12)	-0.01 (0.08)
Single Ascending Dose	0.16 (0.25)	-0.43 (0.25)	-1.05 (0.36)	0.41 (0.25)
Multiple Ascending Dose	-0.09 (0.23)	-1.94 (0.22)	-2.14 (0.31)	-0.17 (0.22)

Table 4.8: Regression Constant

	α_{-1} (SE)	α_0 (SE)	α_{0-1} (SE)	α_2 (SE)
Constant	0.04 (0.11)	2.83 (0.11)	4.10 (0.16)	0.06 (0.11)

In Tables 4.2 to 4.8, we present the regression coefficients obtained using Eq. (4.1) for the abnormal returns (computed using the French-Fama 5-factor model) for the day prior to the event date (-1), on the day of the event (0 and 0–1), and the day after the event date (2). Rows in each table are the different trial properties, and the columns are the regressed coefficients for different days around the event (-1, 0, $0-1$, 2). These coefficients can be interpreted as the average abnormal returns observed for the sponsor company for a trial outcome with that specific property, controlling for other properties of the trial. The abnormal returns are statistically insignificant prior to and after the event date (day -1 and day 2), while they are significant for properties on the days of the event (day 0 and day 0-1).

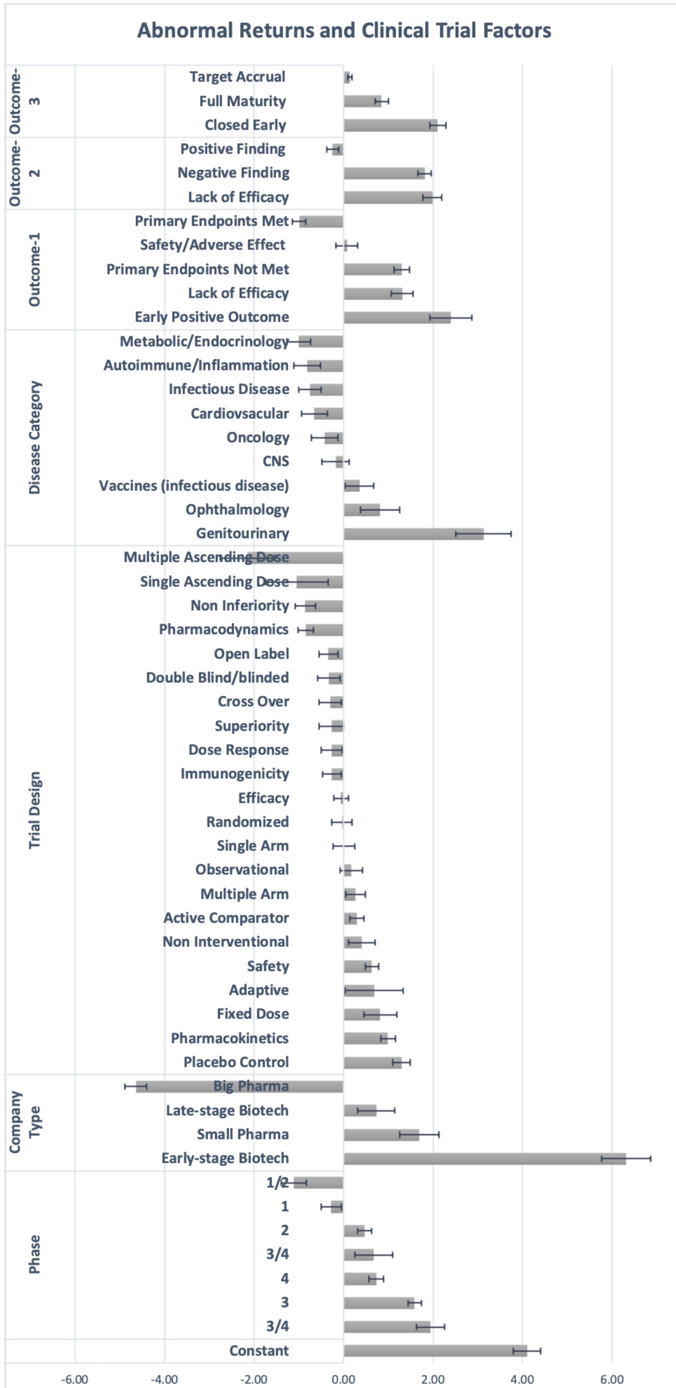


Figure 4-6: Relative abnormal returns due to different clinical trial and sponsor company factors on the day of event (day 0-1).
These abnormal returns are obtained from the regression coefficients using Eq. (4.1) on the day of event (day 0 and day 1). The x-axis is in percentage points and the 95% confidence intervals are plotted on the bar plots.

4.6 Robustness Checks

We perform several robustness checks to determine the sensitivity of our results. In addition to the Fama-French 5-factor model for computing the abnormal returns in the event study, we also apply a constant-mean and market model, which is a single factor model, and the results are in Tables C.3–C.16 in the Appendix. We find that the results obtained using the market and constant-mean model are statistically similar to those using the Fama-French 5-factor model presented in Tables 4.2–4.8.

We also estimate a return attribution model using a subset of our dataset that includes only two outcome categories: when primary endpoints were met (a positive outcome), and when primary endpoints were not met (a negative outcome). These categories use the earliest date of result publication as their event date, and the results are presented in Appendix Tables C.17–C.23. Our findings are consistent with the return attribution model trained on the full dataset.

Standard errors for the abnormal returns due to different factors are computed using 100,000 iterations of the bootstrap process. Each iteration includes sampling the dataset and computing the regression coefficient given by Eq. 4.1. The number of bootstrap iterations is decided by increasing the bootstrap iterations until the average value and standard errors of the regression coefficients converge.

The observations discussed in **Results and Discussion** are based on the statistically significant results of our analysis. As we vary the parameters of the event study such as the length of the estimation and buffer windows (as in Figure 4-2), the results remain statistically similar.

4.7 Conclusion

In this Chapter, we analyzed the impact of clinical trial outcomes on the stock prices of their sponsor companies using one of the largest clinical trials datasets so far constructed, with 13,807 clinical trial outcomes (after filtering) from 2000 to 2020. Given the large size of the dataset and its two-decade span, the findings should represent reasonably accurate measures of the stock-price impact of clinical trial outcomes for sponsor companies.

We identify dates specific to the individual clinical trial's timeline on which the largest movements in the stock's abnormal returns are observed. By fixing the date of that event, we are able to attribute the abnormal returns to clinical trial properties such as outcome type, target accrual, phase, disease therapeutic category, trial design, and sponsor company classification. The attribution of returns to trial properties is performed for abnormal returns on the day of the event and the days after/prior to the event (i.e., days -1, 0, 1, and 2). In the attribution process, we uncover multiple findings about the abnormal returns for various trial properties.

As observed from the significant α in our regression, we find that the clinical trial properties we use cannot completely explain the average abnormal returns observed due to clinical trial outcomes. This implies there are additional factors influencing the company's stock price movements apart from the documented trial properties. One such property may be the "importance" or "buzz" surrounding a clinical trial. For example, the importance of Moderna's mRNA vaccine trial for protection against SARS-CoV-2 was much higher than a clinical trial for any other disease at the same time. Such importance could be captured by analyzing the news and sentiment indexes associated with the trial. In future research, we plan to extend this analysis to quantify the hype or sentiment associated with clinical trials and correlate it with their abnormal returns.

Chapter 5

Finance For Impact: Megafund

Megafund is a structure in which a large number of projects are securitized into a single holding company funded through various debt and equity tranches. In this chapter, we propose an extension of “megafund" to incorporate drug assets from different stages in drug development process:- preclinical, phase 1,2,3 and royalty assets from eight therapeutic areas. Along with the drug development, we utilize the megafund structure to support the fusion energy research. We show that our well diversified megafund can potentially lead to the successful funding for the development of drugs and fusion companies.

5.1 Drug Development

Traditional financing models have struggled to support drug development. Drug development is an expensive and a long process. It requires years of research with extensive amount of funds for development and approval of the new treatment and bears a high risk of failure. Hence, it fails to attract investors and there is a decline in R&D productivity in the biopharma sector in recent years.

Existing work [109] proposed a portfolio approach for raising investments for drug-development process. It proposes investment in multiple drug development projects at a time that diversifies the risk and increases the likelihood of the success. The portfolio is financed using capital raised from issuing debt and equity. It [109] showed the feasibility of the portfolio based approach and found the average investment returns of 8.9-11.4% for the equity holders and 5-8% for the research backed obligation holder for the megafund of $5-15 billion .

Recently, [184] modified the portfolio based megafund framework to make it more realistic by relaxing some of the assumptions (independence of clinical trial projects) and using more accurate parameters (cost, duration and probability of success of clinical trials). It is also used to show the feasability of drug development in multiple areas : orphan drugs [101], alzeihmer's disease [180], pediatric oncology [79], ovarian cancer [57], brain cancer [258], and opiods use disorder [257]. However, most of the studies focussed on one therapeutic area and majority of them acquired drug development assets in portfolios in pre-clinical trial phase.

"Not putting your eggs in one basket" is a widely known phrase in the portfolio management that balances risk and reward in the portfolio. In this chapter, we follow the same and extend the existing megafund framework of [184] to include assets from various stages in the drug development process. We include the assets from preclinical, phase 1,2,3 and royalty assets. The most risky:-preclinical phase is least likely to succeed (around 10-15% probability of success) but is the source of significant returns on successful approval and the least risky:-royalty investment is the source of less risky cash flows. We also diversify the drug assets over eight therapeutic areas:- oncology, metabolic/endocrinology, cardiovascular, central nervous system, autoimmune/inflammation, genitourinary, infectious disease and opthalmology. We also optimise the investment across asset categories to maximise the sharpe ratio.

Apart from the diversification aspect for the investors, the proposed megafund framework generates capital for conducting clinical trials across all disease categories and trial phases. This can lead to development of the successful treatments leading to the decrease in the burden of diseases across all therapeutic categories.

From our simulation of the megafund portfolio, we find that annualized sharpe ratio of 0.41-0.43 and expected annualized returns on equity of 11.3-11.6% for equity investors. We also find that the megafund investment leads to the approval of 19-20 drugs across therapeutic areas. This shows the feasibility of the diversified megafund based approach for the drug development and a viable investment for investors.

5.1.1 Method

Megafund is a portfolio consisting of arbitrary number of drug assets with varying risk, reward characteristics. It acts as a security for investors that is backed by drug assets. For our study, we adopt a framework similar to prior studies that used megafund approach to analyze the performance of alzeihmer's disease [180], pediatric oncology [79], ovarian cancer [57], brain cancer [258] and opiods use disorder[257]. The framework is summarized in the Figure 5-1.

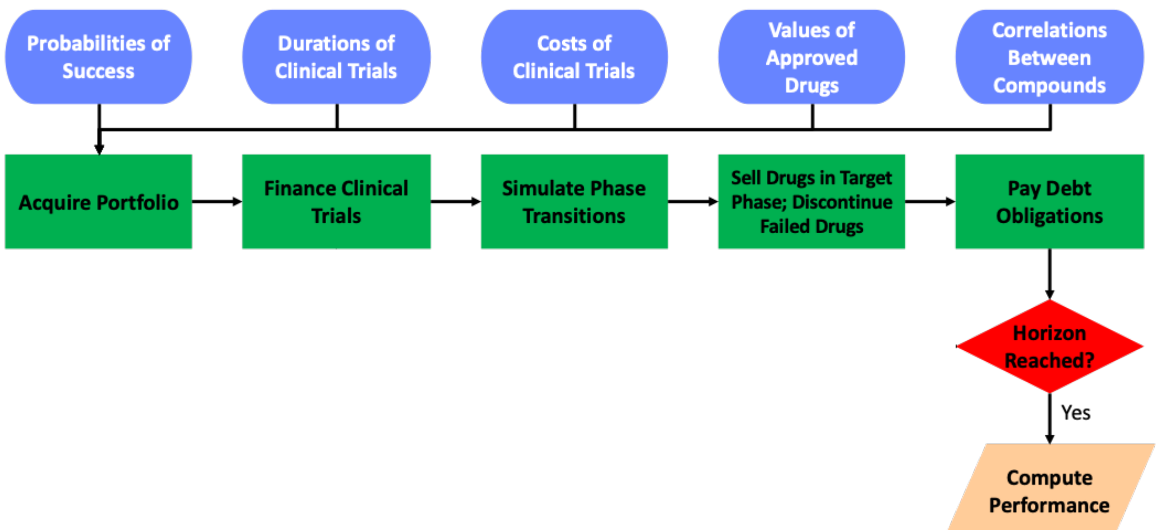


Figure 5-1: Flowchart depicting the megafund simulation framework. The boxes in blue are various parameters estimated/inputs. The boxes in green depict steps of simulation.

The simulation framework is based on the licensing framework commonly used in the bio-pharmaceutical industry [109]. The boxes in the green (Figure 5-1) are the different steps involved in the simulation. First step of portfolio acquisition involves acquiring a majority stake in the drug development process for an upfront payment. We extend the existing framework in the following way, the existing megafund acquires the projects in pre-clinical phase and each project progresses to subsequent clinical trial phases until its approved. In our study we acquire projects that may belong to different stages in the pipeline. The projects includes royalties and drugs at different stages in the approval pipeline (pre-clinical, phase 1, 2 and 3). The drug development projects proceeds to their next steps if it is successful or it is discontinued. Once the project reaches the target stage, it is sold for revenue[1]. Different stages of the projects requires funding for clinical trial expenses and they are financed by megafund. The selling of the successful drugs and royalty investments provides the cash flows for paying debt obligations. Once the fund reaches the horizon, it is liquidated and cash generated is used for fulfilling its debt obligation and equity payment and performance of fund is computed. If the fund is unable to pay the bond payments at any point, it is liquidated. We also optimize investment in different projects maximising the sharpe ratio.

We used multi-period monte carlo simulation to evaluate the performance of the megafund. The simulation to megafund requires different parameters. These parameters are mentioned in blue boxes in Figure 5-1 and they are probability of success, duration of clinical trials, values of approved drugs and correlation between different projects. The probability of success are estimated from [228], [138], [274], [288] and presented in the Appendix Table D.1. The cost of the clinical trials are obtained from [251], [215] and estimates used in the simulation are presented in Table D.2 in Appendix. Similarly the duration of the clinical trials are obtained from [288] and [287] with values presented in Table D.3. The correlation between the outcomes of the drugs belonging to same therapeutic class was assumed to have moderate values of 20% (the drugs treating disease in same therapeutic areas should be more similar) and 5% (low correlation) between the drugs belonging to different therapeutic areas.

The royalty assets were not part of the existing megafund studies. Hence, we modelled the cash flows for the royalty assets using the discounted cash flow methodology described in the article[2]. The details of the cashflow from royalty are available in the Appendix Figure D-2. Each royalty asset is acquired for the cost of 200 million$ and

[1] We assume a sale period of 1 year after approval of drug

[2] https://www.investopedia.com/articles/stocks/06/biotechvaluation.asp. Accessed on 9 September 2022

generates the cash flows for 10 years [3]. Also, the value of drug development projects in different phase needs to be estimated. The estimates of the approved drug in each therapeutic area is obtained from the market research reports. From the approved drug, we estimate the value of drug projects in the earlier stages of development using the following equation:

$$v_x = max(\frac{p_x v_{x+1}}{(1+r_x)^{t_x}} - c_x, 0) \tag{5.1}$$

where v is the risk adjusted net present value, p is the probability of success, r is discount rate, t is the duration of clinical trial, c is the cost of clinical trial in stage of development x. The estimates of the market value of drugs are presented in the Table D.4 in the Appendix.

5.1.2 Results

We construct two megafund portfolios. First is the benchmark portfolio: an equal investment portfolio across across different phases and approximate equal expected investment in each therapeutic area in each phase (Figure 5-2). It involves investing $\approx$ \$2.4 billion in each phase (preclinical, phase 1, 2, 3) and \$2.4B in twelve royalty assets. The total investment is $\approx$ \$13 billion. We construct another portfolio where the \$ amount invested is determined by a greedy algorithm to maximise sharpe ratio of the portfolio. It optimizes for the the number of projects across therapeutic area one phase at a time keeping other parameters constant[4]. Once the number and \$ amount invested across phase and therapeutic area is determined, the megafund is simulated with debt financing. The portfolio composition are summarized in the figure 5-2. We can find that expected investment across the stages of development are around \$2.4B with almost equal investment across therapeutic categories for benchmark. In the optimized portfolio, the investment across different therapeutic categories varies significantly, but the total investment in each phase is $\approx$ \$ 2.4 billion. The portfolio contains a large number of pre-clinical projects due to their low cost while less number of phase 3 projects.

[3]Drug patents usually last for 10 years

[4]assuming all equity financing and equal investment across phases

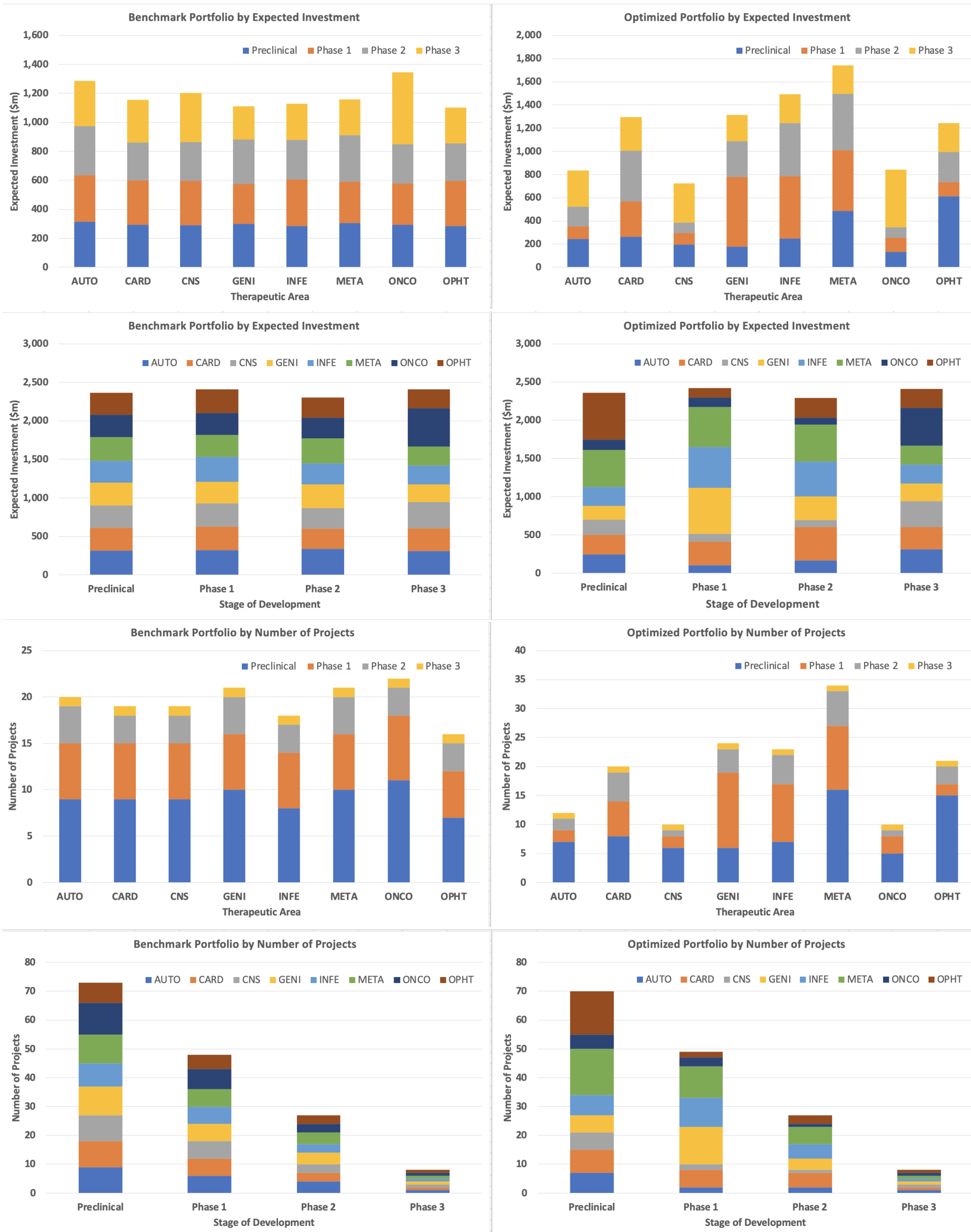


Figure 5-2: Number of projects and investment across phase and therapeutic area for benchmark and optimized portfolio.

The total investment across each phase is $2.4 billion. In benchmark portfolio, the invested amount in each therapeutic area is similar while it varies significantly for optimized portfolio. It is logical that the portfolio contains a large number of pre-clinical projects due to their low cost while less number of phase 3 projects.

In Table 5.1, we include the performance statistics for the megafund. We compute performance statistics like probability of default and expected loss for debt obligations (packaged as junior tranche and senior tranche). In our simulation we assumed a leverage ratio of 80:20 between debt and equity. We find that probability of default for senior tranche is very low (0.4 bps for benchmark portfolio and 2.6 bps for the optimized portfolio). For junior tranche, the probability of default is 15.4 bps and 20bps for benchmark and optimized portfolio respectively. The default rate for senior tranche based on benchmark portfolio are comparable to AAA and AA rated bond.

For the equity investors, we computed multiple performance measures: expected cumulative returns on equity (ROE), standard deviation of cumulative ROE, annualized sharpe ratio, geometric, arithmetic mean and standard deviation of annualized returns on equity, probability of loss, probability of wipeout and probability of annualized returns on equity >10%. All these measures allows us to evaluate the performance of portfolio robustly. From Table 5.1, we observe that sharpe ratio is 0.41 and 0.43 for benchmark and optimized portfolio. The geometric expected annualized ROE is between 11-12%. We find that probability of wipeout, when investors don't receive anything is small (0.2%), while around 70% of time the annualized ROE are greater than 10%.

Although, the ROE are comparable to $S\&P500$ historical annualized returns, the megafund investment led to the multiple successful drug development projects. In Table 5.1, we include the average number of drugs approved in different therapeutic areas. We find that on average 19-20 projects being approved combined from different therapeutic areas with more number of drugs approved in therapeutic areas of genitourinary, infectious disease and metabolic/endocrinology.

As mentioned earlier, we constructed two kind of megafund portfolios- benchmark: equal investment portfolio and optimized: investment based on greedy search to optimize for sharpe ratio. In our simulations, the performance statistics like sharpe ratio (0.43, optimized vs 0.41, benchmark) and number of approved projects (19.9, optimized vs 18.6, benchmark) are slightly better for the optimized portfolio. It is expected as the parameters (number of projects to acquire) are tuned using a greedy

	Benchmark	Optimized
Capital Structure		
Total ($ millions)	12,800	12,800
Senior Debt ($ millions)	8,960	8,960
Junior Debt ($ millions)	1,280	1,280
Equity ($ millions)	2,560	2,560
Tenor (years)	14.5	14.5
Senior		
Prob. of Default (bp)	0.4	2.6
E[Loss] (bp)	0.0	0.1
Junior		
Prob. of Default (bp)	15.4	20
E[Loss] (bp)	5.2	8.3
Equity		
E[Cum. ROE] (%)	483.6	456.4
SD[Cum. ROE] (%)	301.3	270.5
Ann. Sharpe Ratio	0.41	0.43
Geometric Ann. E[Cum. ROE] (%)	12.9	12.6
Arithmetic E[Ann. ROE] (%)	33.4	31.5
Arithmetic SD[Ann. ROE] (%)	20.7	18.5
Geometric E[Ann. ROE] (%)	11.6	11.3
Geometric SD[Ann. ROE] (%)	6.5	6.7
Prob. of Wipeout (%)	0.2	0.2
Prob. of Loss (%)	2.5	2.3
Prob. of Ann. ROE > 10% (%)	70.2	69.3

(a) Performance measures of the megafund portfolio for equity and debt investors obtained from monte-carlo simulations.

	Benchmark	Optimized
Expected Approvals		
Oncology	1.1	0.6
Metabolic/Endocrinology	3.0	4.4
Cardiovascular	1.9	1.8
CNS	1.3	0.8
Autoimmune/Inflammation	2.1	1.3
Genitourinary	3.5	4.1
Infectious Disease	3.2	4.0
Ophthalmology	2.6	3.1
Total	18.6	19.9

(b) The number of approved projects across different therapeutic areas for benchmark and optimized portfolios

Table 5.1

search for the optimized portfolio. However, these parameters may not be optimal and can potentially be improved using a better optimisation methods [5]. Equal weighted portfolio is also a well diversified portfolio across therapeutic areas and clinical trial phases. This maybe a likely reason for its good risk-adjusted performance.

The portfolio construction and performance evaluation involved multiple parameters, we performed sensitivity analysis and investigated the varying performance of megafund on varying the capital structure: from 80:20 debt-equity financing to all equity financing, varying probability of success of the phase transition and changing market value of the drugs.

Leverage

One of the design choice of the megafund construction is the leverage ratio in the raising funds. The portfolio can be financed with debt-equity financing with different leverage. In the Table 5.1, we simulated the portfolio with a debt-equity financing (ratio leverage 80:20). We vary this leverage ratio to evaluate the expected performance of portfolio. We find that high leverage ratio increases the probability of default on the debt. It is logical given that more cash is required for making continuous coupon payment. The probability of default can be reduced by decreasing the leverage, but it comes at the cost of lower return on equity. The sharpe ratio is maximum for the all equity financed megafund for both benchmark (0.45) and optimized (0.48) portfolios. The expected number of drugs approved are maximum for the all equity financed megafund (benchmark: 23.5 and optimized: 25.3). The detailed performance measures are included in the Table D.5 in Appendix.

Probability of Successful Phase Transition

The probability of success is another important parameter in determining the successful phase transition of a drug from pre-clinical stage development to phase 1,2,3 and finally to approval. We vary the probability of the success to observe the effect on the performance. The PoS values used in the study are baseline PoS values obtained

[5] Brute force optimisation will give best possible parameters but it is intractable

on the basis of historical data and literature. We hypothesize that experts involved in portfolio creation can select the drug development projects that can perform better than baseline values. We simulated all equity financed megafund varying the PoS from given in Table D.1 to 1.05, 1.10,1.15, 1.20 and 1.25 times the values in Table D.1. The detailed results are included in Appendix Table D.6. As expected, increasing the PoS leads to more number of approved drugs, from 23.5 to 52.9 (benchmark portfolio) and 25.3 to 56.9 (optimized portfolio)[6]. Higher PoS also leads to higher return of equity and sharpe ratio (increased from 0.42 to 0.45, benchmark portfolio and increased from 0.44 to 0.48, optimized portfolio).

Market Value of Drugs

Market value of drugs is another important factor in our megafund framework. Market value of drugs decide the cost of the assets and revenue if assets are sold. We change the value of approved market drugs to two different set of values: moderate (minor modification in the market value) and aggressive (large modifications in the market value) given in the Table D.7. We also modify the discount factor in the Equation 5.1 that changes the value of the pre-clinical and phase 1, 2, 3 approved drugs. We don't observe a significant change in the megafund performance on varying these factors (refer Table D.8). There is a increase in the default probability for senior tranche bonds for benchmark portfolio (increased to 1.2 bps from 0.4 bps) while the returns on the equity increased to 12.5-6% from 11.3-6%. Also, we find slightly more number of drugs approved (increased to 19-22 from 18-20). Hence, the megafund performance is robust to change in the market value of drugs.

5.2 Fusion Energy Research

The case for investing in fusion energy has never been greater, given increasing global energy demand, high annual carbon dioxide output, and technological limitations for wind and solar power. Nevertheless, financing for fusion companies through tra-

[6]The PoS is varied for all equity financed megafund

ditional means has proven challenging. While fusion startups have an unparalleled upside, their high upfront costs, lengthy delay in payoff, and high risk of commercial failure have historically restricted funding interest to a niche set of investors. Drawing on insights from investor interviews and case studies of public–private partnerships, we propose the utilization of a megafund structure in which a large number of projects are securitized into a single holding company funded through various debt and equity tranches, with first loss capital guarantees from governments and philanthropic partners. The megafund exploits many of the core properties of the fusion industry: the diversity of approaches to engender fusion reactions, the ability to create revenue-generating divestitures in related fields, and the breadth of auxiliary technologies needed to support a functioning power plant. The model expands the pool of available capital by creating tranches with different risk–return tradeoffs, and providing a diversified "fusion index" that can be viewed as a long hedge against fossil fuels.

5.2.1 Method

The summary of steps of megafund are included in the Figure 5-3. We used multi-period monte carlo simulation to evaluate the performance of the megafund. The boxes in the blue are the parameters that are input for the simulation. The parameters are probability of success of successful phase transition, length, cost of each phase, spinoff probabilities, correlation between projects and value of projects (cash that can be generated by selling a project). These parameters are estimated via the consultation with the industry experts. First step of portfolio acquisition involves acquiring a fixed percentage stake in the projects. Different stages of the projects require funding and they are financed by the megafund proportional to the ownership. The selling of the successful projects generates the cash flows for paying debt obligations. Once the fund reaches the horizon, it is liquidated and cash generated is used for fulfilling its debt obligation, equity payment and performance of fund is computed. If the fund is unable to pay the bond coupon payments at any point, it is liquidated.

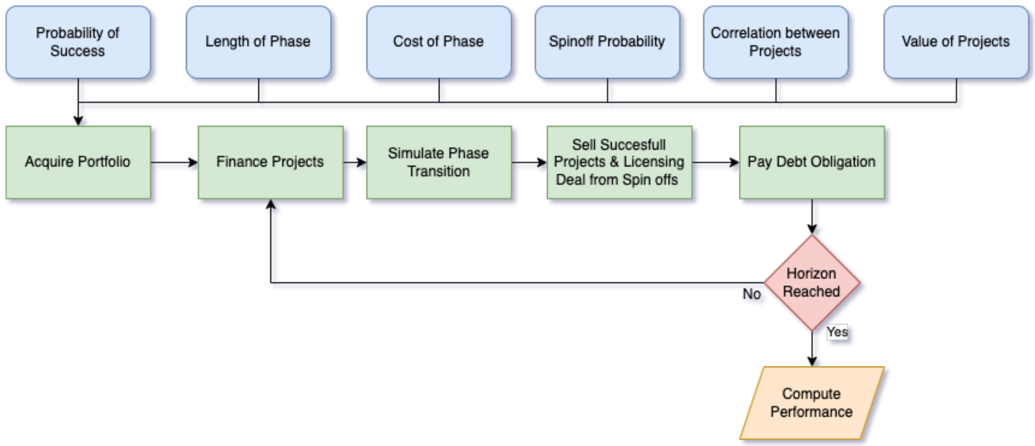


Figure 5-3: Flowchart depicting the megafund simulation framework. The boxes in blue are various parameters estimated/inputs. The boxes in green depict steps of simulation

Results

We construct megafund portfolio by acquiring 10% stake in each company (total 18). The tenor of the fund is 20 years. Whenever the capital is raised via debt, we have three tranches: junior (interest 6%), mezzanine (interest 4%) and senior (interest 2%) along with equity. In Table 5.2, we include the performance statistics of the megafund. The total capital raised is \$3 Billion using equity and debt in the ratio of 100:0, 60:40 and 40:60. We find annualized ROE is in range of -0.78-2% and annualized sharpe ratio in ranges from -0.02 to 0.11. When higher percentage of capital is raised via debt, we find scenarios when megafund defaults more often. For debt-to-equity ratio of 40:60, we find low probability of default for the junior tranche (0.01%) and no default scenarios for mezzanine and senior tranche. For larger capital raised using debt, we observe higher default probabilities because less cash available for coupon payments. We find that the number of successful projects is few due to the longer duration of the projects. At the end of megafund tenor, we find 4 projects in breakeven and 3 projects in commercialization stage.

In our modelling, when the project successfully finishes R&D stage, there is a potential of spinoff companies using the new technology developed during research. We include the number of successful spinoff projects from our portfolio in Table 5.3.

Measures	100:0 (Equity:Debt)	60:40 (Equity:Debt)	40:60 (Equity: Debt)
ROE (Total)	39.12	11.85	-14.50
ROE (Arithmetic Avg)	1.96	0.59	-0.73
ROE (Geomeric Avg)	1.66	0.56	-0.78
Pr(ROE<-10%)	3.36	81.16	87.79
Sharpe (Annual)	0.11	0.02	-0.02
# Prototyping	0	0	0.1
# Breakeven	4.2	4.3	4.3
# Commercialization	3.4	3.1	2.7
# Finished	0.1	0.1	0.1
Junior Default(%)	0.0	0.01	38.12
Mezzanine Default(%)	0.0	0.0	2.79
Senior Default(%)	0.0	0.0	0.01

Table 5.2: Performance statistics of the megafund.
The ratio of equity-to-debt is varied from 100:0, 60:40 and 40:60.

We find 4 laser technology, 5 medical scanning and 6 magnet technology company spinoffs from our portfolio. Later, we will find that cashflows from the licensing deal contributes significantly to the portfolio performance.

Sensitivity Analysis

As discussed earlier, the simulation requires multiple input parameters. These parameters are cost/funding required for each stage, duration of stages, probability of successful phase transition, revenue obtained from selling of the project at different stages and revenue from the licensing deal for the spinoff companies. In Table 5.4, we present the performance of equity funded megafund varying the value of successful project from $66 Billion to $44 and $22 Billion. As expected, we find decreasing ROE

Spin off Companies	Average Number	Standard Deviation
Laser	4.0	1.8
Magnets	6.0	1.0
HTS	0.9	0.3
Cyclotron	3.0	1.0
EM Aircraft Launch	2.3	1.6
Thermal Measurement	2.3	1.6
Robotics	1.7	1.8
Medical Scanning	5.5	1.9
Life Science	1.7	0.8
Material Science	1.7	0.8

Table 5.3: Average number of spin-off companies from the successful megafund.

by decreasing the value of successful project. We also observe that lowering the value of licensing deal from \$300mil to \$30mil (keeping value of successful project fixed to \$22 Billion) decreased the average annual ROE from negative 1.13% to negative 2.96%. Hence, along with the revenue generated from the successful projects, the revenue from spinoff licensing deals an important component in generating cash flow and returns for investors.

In Table 5.5, we include the performance of equity funded megafund varying the parameters like probability of successful phase transition, funding of projects in different phases. As expected, increasing the probability of success leads to higher returns on equity because of fewer number of projects failing. Similarly, lower costs in each phase have higher returns on equity. We observe higher ROE on both decreasing and increasing the duration of each stage. For lower duration, a greater number of projects reaches the final stages in the fixed tenor.

	Revenue:$66bil	Revenue:$44bil	Revenue: $22bil	Revenue:$22bil
Measures	Spinoff: $300mil	Spinoff: $300mil	Spinoff: $300mil	Spinoff: $30mil
ROE(Total)	39.12	9.39	-20.34	-45.2
ROE(Arithmetic Avg)	1.96	0.47	-1.02	-2.26
ROE(Geometric Avg)	1.66	0.45	-1.13	-2.96
P(ROE<-10%)	3.36	42.14	78.05	88.62
Sharpe (Annual)	0.11	0.04	-0.15	-0.4
# Prototyping	0	0	0	0
# Breakeven	4.2	4.2	4.2	4.4
# Commercialization	3.4	3.4	3.4	3.2
# Finished	0.1	0.1	0.1	0.1

Table 5.4: Megafund Performance.
We vary the value of the successful project (revenue generated after selling a successful project) from $66bil to $44 and $22bil. We also vary the cash flow generated from licensing deal of the spin off companies from $300mil to $30mil. Megafund is equity funded.

5.3 Conclusion

For drug development, we extended the existing megafund framework to incorporate drug projects in various stages of development (pre-clinical, phase 1,2,3 and royalties) and from eight different therapeutic areas and showed the feasibility of the megafund. Along with generating 11.3% annualized returns on equity, it also leads to the development of 19-20 drugs that can potentially be used to decrease the burden of diseases. We robustly evaluate the megafund by computing multiple performance measures and conducted sensitivity analysis on the capital structure (leverage ratio), probability of successful phase transition and market value of drugs.

For drug development, the diversified megafund approach can be a source of funds for drug projects that gets abandoned in the drug development process due to lack of funding and it can potentially lead to the development of successful treatment. From the investment perspective, megafund can a investment opportunity where the investors are trying to diversify away from the traditional assets. We found that

	Baseline	PoS	PoS	Cost	Cost	Length	Length
	1x	0.8x	1.2x	0.8x	1.2x	0.8x	1.2x
ROE(Total)	39.12	21.52	56.41	54.97	23.5	61.86	44.54
ROE(Arithmetic Avg)	1.96	1.08	2.82	2.75	1.17	3.09	2.23
ROE(Geometric Avg)	1.66	0.98	2.26	2.21	1.06	2.44	1.86
P(ROE<-10%)	3.36	3.62	2.69	0.07	30.55	59.59	0.81
Sharpe (Annual)	0.11	0.09	0.12	0.15	0.29	0.09	0.1
# Prototyping	0	0	0	0	0	0	2
# Breakeven	4.2	2.7	4.9	4.2	4.3	1.1	5.8
# Commercialization	3.4	1.7	4.9	3.4	3.2	1.4	1.7
# Finished	0.1	0.1	0.2	0.1	0.1	0.5	0.1

Table 5.5: Megafund Sensitivity Analysis.
Megafund portfolio performance varying the parameters like probability of successful phase transition, cost of each stage and length of each stage of the project. The baseline equity funded megafund performance are included in column 2. For other parameters, we vary the estimates to 0.8xbaseline and 1.2xbaseline keeping other parameters constant. The value of successful project was fixed to \$66bil and spinoff values were fixed to \$300mil..

along with high returns on equity, the senior tranche issued on the megafund had the probability of default of 0.4 basis points which is comparable to the AAA and AA rated bonds.

The megafund included modelling multiple parameters:- probability of success, cost & duration of clinical trials, value of approved drugs, correlation between compounds and cash flows from royalty assets. We modelled and adapted them from best possible sources in literature but the assumptions and parameters can be modified and performance can be evaluated using the open source software.

For fusion research, we find positive returns on equity (ROE) and low default rates for the capital raised using debt. However, the ROE and default rates depends on the value of the successful/in-pipeline projects and cashflows from the licensing deals of the spinoff companies. We performed a sensitivity analysis to show the varying ROE for the estimates of the value of pipeline projects, spinoff companies licensing deals and other parameters.

Chapter 6

Finance for Impact: Quantifying the Returns of ESG Investing

In this Chapter, we quantify the financial performance of environmental, social, and governance (ESG) portfolios in the U.S., Europe, and Japan, based on data from six major ESG rating agencies. We document statistically significant excess returns in ESG portfolios from 2014 to 2020 in the U.S. and Japan. We propose several statistical and voting-based methods to aggregate individual ESG ratings. We find that aggregate ESG ratings improves portfolio performance. In addition, we find that a portfolio based on Treynor-Black weights further improves the performance of ESG portfolios. Overall, these results suggest there is a significant signal in ESG rating scores that can be used for portfolio construction despite their noisy nature.

6.1 Introduction

In recent years, environmental, social and governance (ESG) concerns have gained considerable attention from investors. The market for ESG investing is currently estimated at $9 trillion in the U.S.,[1] while the number of organizations that are signatories of the United Nations Principles for Responsible Investment (UN-PRI) has increased from 450 in 2016 to 4935 in 2022, representing over $100 trillion of

[1] https://www.rockpa.org/guide/impact-investing-introduction/

assets under management.

Investors who want to implement ESG factors in their portfolios typically rely on ESG scores provided by third-party rating agencies that specialize in measuring ESG performance. There is a rapidly growing number of such rating agencies, and in our sample, we rely on ratings from some of the largest agencies, such as MSCI Inc, S&P Global, ISS, Moody's ESG Solutions, Reprisk, and TruValue Labs.Each rating agency has a proprietary methodology for the calculation of their ratings. This process typically involves gathering data from a variety of sources, including yearly regulatory filings, media reports, and self-disclosed data from firms and international organizations. The choice of different data methodologies and sources can lead to a substantial divergence between rating providers [27].[2] ESG ratings, and sustainable investing in general, have received a number of high-profile critiques. For example, the *Economist* dedicated a recent cover story to ESG investing, concluding that ESG ratings are too complex and contain too much measurement error to be useful. This raises the question whether ESG ratings are in fact useful for portfolio construction, and if so, how to optimally exploit the signal in ESG ratings, despite their noisy nature.

One of the most important critiques of ESG ratings raised by regulators concerns the fiduciary responsibility of financial institutions. Portfolios constructed using the ESG scores are constrained, and therefore, if such a constraint reduces portfolio returns, it could be considered an illegal act — especially if the financial institution fails to clearly inform the investor of such a possibility. In general, most of the regulatory commentary is based on the intuition that restricted portfolios are by necessity less profitable than unrestricted ones. However, this intuition is correct only if a constraint is orthogonal to returns. That is not the case if the selection mechanism is associated with the fundamental characteristics of the stocks. Consequently, understanding if ESG scores are associated with excess returns is of crucial importance to investors, regulators, financial institutions, and ultimately, to those that care about ESG impact to society in general.

[2] Furthermore, [26] show that these ratings contain a considerable amount of measurement error.

In our empirical analysis, we construct ESG portfolios for the U.S., European, and Japanese stock markets, using ESG scores from six major rating agencies. We quantify the excess returns of these portfolios with respect to standard asset pricing models, including the capital asset pricing model (CAPM) and several Fama-French factor models. Our sample ranges from 2014 to 2020. We find a wide range of excess returns in portfolios constructed using different ESG scores. For example, the MSCI-based portfolio that goes long the top quartile of stocks with the highest ESG ratings and short the bottom quartile achieves a statistically significant annual alpha of 3.8% in excess of the Fama-French five-factor model in the U.S., while the same portfolio using other ESG ratings shows much lower (and usually neutral) excess returns. In addition, the same rating agency may have very different excess returns across regions. This instability in coefficients should have been expected, considering the sizeable noise in ESG scores.

To address the problem of noise, we propose several different ways to aggregate ESG scores across vendors. We construct a single measure of ESG by combining individual ESG scores from six different vendors using various statistical and voting aggregation techniques, including simple averages, the Mahalanobis distance, principal component analysis, average voting, and singular transferable voting. Our goal is to retain the ESG signal in the aggregated rating while attenuating the noise. Different aggregation methods will necessarily weight the ESG scores from rating agencies differently. For example, the simple average attributes equal weights to scores from all vendors, while the Mahalanobis distance aggregates ratings based on their variance-covariance, and principal component analysis weights the rating agencies in such a way to retain the direction of their maximum observed variance.

We find that aggregating individual ESG ratings improves portfolio performance significantly. We construct sorted ESG portfolios (from high to low scores) and analyze their risk-adjusted returns, excess returns, and exposures to fundamental factors. In particular, we find that portfolios in the U.S. based on the Mahalanobis distance achieve the highest annualized alpha, over 6%, while portfolios based on singular transferable voting achieve the highest annualized alpha, over 6% in Europe and 9%

in Japan.

The empirical evidence on excess returns of ESG investing has been mixed in the existing literature. Some document a positive relationship between ESG scores and excess returns (see, for example, [96], [154], [174], and [10]), while others find a negative relationship (see, for example, [58], [98], and [34]). [26] describe a theoretical model explaining both relationships. They explain the positive realized returns by unexpected inflows into stocks with high ESG performance. As these inflows level out, the expected returns become lower. Put differently, high ESG firms benefit from a lower cost of capital due to higher market capitalization. Another possible explanation is found in omitted variable bias, in particular the omission of management quality. If good ESG performance is correlated to high management quality, the link between ESG performance and returns would no longer be causal. In our sample from 2014 to 2020, we found a positive relationship in the U.S. and Japan most likely due to inflows of funds from new ESG investors into high ESG stocks. For example, [25] find that MSCI rating changes drive changes in ESG mutual fund holdings in the U.S. market, albeit with a very slow integration of up to 18 months. They also show that this correlates temporally with returns.

We find that the portfolio construction methodology proposed by [186] further improves the performance of ESG portfolios. [186] methodology begins by quantifying the excess returns for individual assets using a small number of parameters,[3] then applies Treynor-Black weights to optimize the Sharpe ratio of an ESG portfolio, in which the weights are proportional to the rank of the ESG score of each firm.[4] Using this framework, we achieve improved excess returns in ESG portfolios, especially for portfolios with a large number of assets. This is valid in particular for portfolios constructed to go long the top 4 deciles of stocks with the highest ESG ratings and short the bottom 4 deciles of stocks with the lowest ratings, so that weights based on the rank of each firm's ESG score have a meaningful impact.

[3] Specifically, it uses the cross-sectional correlation between ESG scores and excess returns of each stock.

[4] We compare the excess returns of ESG portfolios using the model of [186] to their forward-looking realized excess returns and find a high degree of consistency between the two, thereby validating the model.

Some investors may prefer to rely on E, S, or G scores individually in the creation of their portfolios. Consequently, we also investigate the aggregation of individual E, S, and G scores across vendors, and analyze the excess returns of top-bottom sorted portfolios. We find the highest excess returns for portfolios based on E scores in the U.S. and Japan. In portfolios based on S and G scores, we find positive excess returns only for some portfolios and aggregation methods.

This chapter is related to several strands in the literature. The first strand is about disagreement between ESG providers. Our work is based on the growing literature that highlights the divergence between ESG ratings (see, for example, [91], [250], [27], and [121]).

Second, our work is also closely related to literature that explores the relationship between ESG and stock returns. Some research shows higher returns [96, 154, 174], while other work shows a negative relationship both empirically [34] and theoretically [300]. [227] show that the high returns for green assets in recent years reflect unexpectedly strong increases in environmental concerns, not high expected returns. Our work differs in our acknowledgement of the noisiness of ESG ratings and our proposal of different aggregation methods.

Third, our research is related to the nascent literature in dealing with measurement noise in ESG ratings and its impact on returns. [26] use instrumented variable regressions to remove the noise in one version of the ESG score using others. In our work, we improve the signal using aggregation methods that combine multiple sources of data, and we leverage the optimal portfolios of [186] to further improve the performance of these ESG portfolios.

In the remainder of this Chapter, Section 6.2 discusses the methodology used to quantify excess returns, construct portfolios, and aggregate individual ESG scores. We then discuss the data and present summary statistics in Section 6.3. Section 6.4 presents our extensive empirical analysis using different ESG scores and portfolio construction methodologies. Finally, we conclude in Section 6.6.

6.2 Methodology

In this section, we describe the methodology used to construct portfolios based on ESG scores. We discuss our strategy to quantify excess returns for individual stocks, the portfolio construction methodologies, and several methods to aggregate multiple ESG scores.

6.2.1 Quantifying Excess Returns

We start with describing a methodology first proposed by [186], which we adapt to ESG portfolios. We quantify the excess returns (alphas) of individual stocks ranked by their ESG scores. This allows us to optimize the weights used in ESG portfolios and to quantify their portfolio returns.

We consider a universe of N stocks with returns R_{it} that satisfy the following linear multi-factor model (e.g., the Fama-French factor model):

$$R_{it} - R_{ft} \;=\; \alpha_i \;+\; \beta_{i1}\left(\Lambda_{1t} - R_{ft}\right) \;+\; \cdots \;+\; \beta_{iK}\left(\Lambda_{Kt} - R_{ft}\right) \;+\; \epsilon_{it} \tag{6.1}$$

$$\text{such that} \;\; \mathbb{E}[\epsilon_{it}|\Lambda_{kt}] = 0 \;\;,\;\; k = 1, \ldots, K \tag{6.2}$$

where Λ_{kt} is the k-th factor return, $k = 1, \ldots, K$, R_{ft} is the risk-free rate, α_i and β_{ik} are the excess return and factor betas, respectively, and ϵ_{it} is the idiosyncratic return component.

ESG investors typically rank stocks according to their ESG scores, which we denote by ESG_i, and we use $\alpha_{[i:N]}$ to represent the alpha of the i-th ranked stock.[5] [186] show that the expected values, variances, and covariances of these ranked alphas are

[5] In the statistics literature, these indirectly ranked variables are termed *induced order statistics* [31], because they are ranked not by their own values (α_i in our case) but by the values of another variable (ESG_i in our case). As such, α_i's are modeled as random variables to reflect the fact that they may be correlated with the ESG scores. This specification was used in [181] to represent the cross-sectional estimation errors of intercepts derived from CAPM regressions. In the current context, we interpret the randomness in α_i as a measure of uncertainty regarding the degree of mispricings of stocks, which is similar to the treatment in [226].

given by

$$\mathbb{E}(\alpha_{[i:N]}) = \sigma_\alpha \cdot \rho \cdot \mathbb{E}(Y_{i:N}), \tag{6.3}$$

$$\mathrm{Var}(\alpha_{[i:N]}) = \sigma_\alpha^2 \cdot (1 - \rho^2 + \rho^2 \cdot \mathrm{Var}(Y_{i:N})), \tag{6.4}$$

$$\mathrm{Cov}(\alpha_{[i:N]}, \alpha_{[j:N]}) = \sigma_\alpha^2 \cdot \rho^2 \cdot \mathrm{Cov}(Y_{i:N}, Y_{j:N}), \tag{6.5}$$

for $i, j = 1, 2, \ldots, N$, and $i \neq j$. Here, ρ is the cross-sectional correlation between α_i and ESG_i,[6] σ_α is the standard deviation of α_i, and $Y_{1:N} < Y_{2:N} < \cdots < Y_{N:N}$ are the order statistics of N independent and identically distributed standard Gaussian random variables.

To estimate Equations (6.3)–(6.5) in practice, for each year, we first perform a time series regression[7] using daily returns to obtain an estimate of α_i for each stock. We then compute ρ and σ_α based on ESG_i and the estimated α_i. Finally, we calculate the moments related to $Y_{i:N}$ based on a simulation of N standard normal random variables.

The results in Equations (6.3)–(6.5) are useful for two reasons. First, they allow us to construct optimal Treynor-Black portfolios that maximize the Sharpe ratio of the ESG portfolios. In particular, the expected excess returns in Equation (6.3) are crucial in determining the weights of those portfolios. The advantage of using this specific framework lies in its robustness, because only two parameters, ρ and σ_α, need to be estimated, compared to traditional Markowitz portfolios, which are known to produce unstable weights [41, 276].

Second, the results in Equation (6.3) allow us to quantify the excess return of any ESG portfolio, α_p, with weights ω_i, $i = 1, \ldots, N$:

$$\mathbb{E}\left(\alpha_p\right) = \sum_{i=1}^{N} \omega_i \mathbb{E}(\alpha_{[i:N]}) = \rho \sigma_\alpha \sum_{i=1}^{N} \omega_i \mathbb{E}\left[Y_{i:N}\right]. \tag{6.6}$$

[6] [186] assume that they are jointly normally distributed, and [185] generalize the framework to arbitrary marginal distributions.

[7] The factors in the time series regression are decided on the basis of the Capital Asset Pricing Model or selected Fama-French factor models.

This provides an estimate of the alpha for ESG portfolios, which we refer to as the *model-implied alpha* henceforth. We validate the model-implied alpha empirically against a forward-looking estimate of alpha in Section 6.4. In particular, the realized alpha for any portfolio can be computed by performing a time-series regression based on a multi-factor model, such as the CAPM, the Fama-French three-factor model (FF3), or the Fama-French five-factor model (FF5). At time t, the forward-looking realized one-year alpha is computed using returns $r_{t+1}, r_{t+2}, r_{t+3}, \ldots, r_{t+252}$ for a portfolio p. The realized alpha is represented by α_p^f, where f may be CAPM, FF3 or FF5.

6.2.2 Portfolio Creation

Given a set of ESG scores for all firms, we construct ESG portfolios and estimate their performances. At time t, we sort the stocks based on ESG scores and construct three long/short portfolios. The first, $pf_{(\pm 10)}$, represents a portfolio that goes long the top decile of stocks with equal weights (denoted by $pf_{(+10)}$) and short the bottom decile of stocks with equal weights (denoted by $pf_{(-10)}$). The second, $pf_{(\pm 25)}$, represents a portfolio that goes long the top quartile of stocks with equal weights (denoted by $pf_{(+25)}$) and short the bottom quartile of stocks with equal weights (denoted by $pf_{(-25)}$). The third, $pf_{(\pm 40)}$, represents a portfolio that goes long the top 4 deciles of stocks with equal weights (denoted by $pf_{(+40)}$) and short the bottom 4 deciles of stocks with equal weights (denoted by $pf_{(-40)}$). The total weights of individual stocks on the long and short sides are both set to one to give all portfolios the same amount of leverage. We refer to these three portfolios as the ± 40, ± 25 and ± 10 portfolios henceforth. These portfolios are rebalanced once a year because ESG scores are updated by vendors once a year (with the exception of Reprisk). We observed very high cross-sectional autocorrelations between scores, as shown in Table E.1 in the supplementary material.

In addition to these equal-weighted portfolios, we build optimized Treynor-Black portfolios using the model-implied alphas for individual stocks in Equations (6.3)–(6.5), which use the rank of stocks in the ESG sorted portfolio. For example, in

$pf_{(+10)}$ for equal-weighted portfolios, all stocks in the 10th decile (percentiles 90 to 100) are given equal weights. However, in Treynor-Black weighting, higher-ranked stocks are given larger weights than lower-ranked stocks, if the correlation ρ between the ESG score and stock alpha is positive. Specifically, [186] show that the weight of the ith ranked stock in a universe of N stocks can be approximated by the following equation:[8]

$$\omega_i \propto \Phi^{-1}(\zeta_i) \tag{6.7}$$

where $\zeta_i = i/N$ and Φ^{-1} is the inverse of the cumulative standard normal distribution.

Once these ESG portfolios are constructed, we can combine them further with any other portfolio. The most natural application is to combine the active (ESG) portfolio with a passive index, such as the market portfolio. The returns of the combined portfolio are given by:

$$r_{active+passive} = \omega_A * r_{active} + (1 - \omega_A) * r_{passive} \tag{6.8}$$

where r_{active} can be any return of a top-bottom equal-weighted or Treynor-Black weighted portfolio, $r_{passive}$ is the return of the passive portfolio (e.g., the market index), and ω is the weight of the active portfolio. In our analysis, we have fixed ω_A at 0.5 as an illustrative example.

In Section 6.4, we evaluate the performance of all equal-weighted, Treynor-Black, and active-plus-passive portfolios.

6.2.3 ESG Rating Aggregation

*Berg et.al.*citeberg2021esg show that ESG ratings are certainly noisy but nevertheless contain a signal. The measurement error inherent to ESG ratings makes it difficult to find significant risk premia. To solve this problem, we propose several different ways of aggregating ESG scores. Let $ESG_{i,t,j}$ be the ESG score of company i rated by

[8] The approximation holds when the number of stocks is large and stocks have identical idiosyncratic volatilities.

vendor j at time t. We then compute $ESG_{i,t,m}$, where m is the aggregation method. We describe each aggregation method in the following paragraphs, and analyze the performance of portfolios based on these aggregated ESG scores in Section 6.4.

Equal Weighted Average

The first and simplest aggregation method under consideration is the equal-weighted cross-sectional average of all ESG ratings. The average is a widely used method for noise attenuation, and we intuitively expect the ESG signal to be stronger here than in the case of a single rating. We use AVG to indicate the equal-weighted average rating.

$$ESG_{i,t,avg} = \sum_{j \in \{\text{MSCI,S\&PGlobal,...,TVL}\}} ESG_{i,t,j} \tag{6.9}$$

PCA

Principal Component Analysis (PCA) is a widely used dimensionality reduction method. In some cases, it can also be used as a tool for noise reduction. The PCA performs a change of basis transformation and projects the data in the direction of the maximum variance. If errors are similarly distributed between ratings, it will minimize the information loss. We treat the ESG scores from the six vendors in our sample as high-dimensional data and obtain its lower dimensional (1-d) representation as the aggregate score.

$$ESG_{i,t,PCA} = PCA(ESG_{i,t,MSCI}, ESG_{i,t,S\&PGlobal}...ESG_{i,t,TVL})$$

We use PCA to indicate the aggregate score obtained by principal component analysis.

Mahalanobis Distance

The Mahalanobis distance is the distance between two points after accounting for variances and covariances across dimensions. In brief, highly correlated ESG scores will not be given excess weight in the calculation of the aggregate score. Let the ESG scores of firm i be the vector x and y be the vector with minimum ESG scores. We then calculate the Mahalanobis distance as follows:

$$\begin{aligned} x_{i,t} &= (ESG_{i,t,MSCI}, ESG_{i,t,S\&PGlobal}...ESG_{i,t,TVL}) \\ y_t &= (min_i(ESG_{i,t,MSCI}), min_i(ESG_{i,t,S\&PGlobal})...min_i(ESG_{i,t,TVL})) \\ ESG_{i,t,Maha} &= \sqrt{(x_{i,t} - y_t)^T S_t^{-1} (x_{i,t} - y_t)} \end{aligned} \tag{6.10}$$

where S_t is the variance-covariance matrix for the ESG ratings at time t. We refer to the aggregate score obtained by the computation of the Mahalanobis distance as $MAHA$.

Voting Average

The voting average is based on the theory of social choice, that is, the aggregation or combination of individual preferences in collective decisions. In voting aggregation methods, the ESG scores from a rating agency are considered as a ranked list of choices and the different agencies are considered as voters. In the voting average process, each rank is averaged to compute an aggregate rank of each firm. It is represented by AVG_{vote} in our subsequent analysis.

Voting STV

The Singular Transferable Vote (STV) method is another way of aggregating the ESG ratings. In this aggregation method, the least preferred candidate is eliminated and the vote is transferred to the next preferred candidate. The process is repeated until all the candidates are eliminated, one by one. The eliminated candidates are ranked in the order of elimination to form a ranked list of the candidates. The aggregated ESG score using this method is represented by STV_{vote} in our subsequent analysis.

Optimized ESG Score

The optimized ESG score is a linear combination of ESG scores from all rating agencies. The weights are derived from an optimization that maximizes the cross-sectional correlation between ESG scores and the one-year future excess returns of the stocks.

$$\begin{gathered} ESG_{i,t,opt} = \sum_{r \in MSCI, S\&PGlobal...} w_r * ESG_{i,t,r} \\ w_r : max(avg_t(corr_i(ESG_{i,t,opt}, \alpha_{i,t}))) \end{gathered} \tag{6.11}$$

Since the optimization involves the use of out-of-sample excess returns, the excess returns obtained using $ESG_{i,t,opt}$ cannot be realized. However, the ESG scores optimized in this way have the maximum correlation with excess returns that can be achieved by a linear aggregation of ESG scores from different rating agencies. The aggregate scores obtained using optimization are referred to as OPT in our analysis.

6.3 Data and Summary Statistics

6.3.1 Data

We obtained data from six leading ESG rating providers, including MSCI, S&P Global, ISS, Moody's ESG Solutions, Reprisk and TruValueLabs.[9] We observe non-overlapping coverage in our dataset across rating agencies. Hence, for a fair comparison across different providers, we only include those firms with observations from all six rating agencies. We also have E, S and G ratings in our dataset for all agencies, except for TruValueLabs, which does not offer such scores.

For our analysis, we classify firms into three different regions: the U.S., Europe, and Japan. The total number of firms in each region are 633 in the U.S, 547 in Europe, and 274 in Japan. Our sample spans from March 2014 to 2020. The relatively short time series is explained by the fact that sustainable investing is a fairly recent

[9]We also have data from Sustainlytics and Refinitiv. However, Refinitiv and Sustainalytics have changed their methodologies over time, and their ESG scores were backfilled using these new methodologies [24]. Adding these two providers to our analysis would introduce a forward-looking bias to our analysis.

phenomenon. Since ESG ratings from different providers have different scales, we renormalize them to have zero mean and unit variance in the cross-section.

The daily returns were queried from the Refinitiv workspace[10]. For the calculation of excess returns, we obtained our data from the Fama-French data library[11]. We retrieved the daily MSCI index returns for U.S., Europe and Japan from MSCI Index Solutions[12].

6.3.2 Summary Statistics: Excess Returns and ESG

We averaged the cross-sectional rank[13] correlations between different individual and aggregated ESG scores over time, and present them in Table 6.1. We show that the correlation between different ESG ratings vary over time. Some scores are highly correlated with each other, like ISS, MSCI, SPGlobal and Moody's. Rather surprisingly, Reprisk has a negative correlation with the other ratings. Among the aggregate scores, AVG, PCA and AVG$_{vote}$ exhibit relatively high correlations (>80%). However, the aggregate scores based on STV voting are significantly different from other aggregate scores, given their smaller correlation coefficients (60-70%). We find relatively similar correlations (in the range of 40-50%) between individual ESG scores and scores computed using the Mahalanobis distance (MAHA). This is due to the variance-covariance normalization property of the Mahalanobis distance.

In Table 6.2, we present the cross-sectional mean and standard deviation of excess returns (alphas $\alpha_{i,t}$: the one-year future alpha of the stock i at time t) computed as follows: $mean_\alpha = avg_t(avg_i(\alpha_{i,t}))$, $std_\alpha = avg_t(std_i(\alpha_{i,t}))$, where i is the company and t is the time. We find that the cross-sectional standard deviation of excess returns varies between 18 and 25%, while the cross-sectional mean of excess returns varies widely between regions. These parameters are necessary to quantify the excess returns of portfolios constructed as described in Section 6.2.1.

[10] https://www.refinitiv.com/en/products/refinitiv-workspace
[11] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
[12] https://www.msci.com/our-solutions/indexes
[13] In the construction of a long-short ESG portfolio, the rank of a company matters more than its score value.

	ISS: ESG	MSCI: ESG	Reprisk: ESG	SPGlobal: ESG	TVL: ESG	Moody's: ESG	AVG: ESG	PCA: ESG	MAHA: ESG	AVG_{vote}: ESG	STV_{vote}: ESG
ISS: ESG	1.00	0.40	-0.24	0.59	0.12	0.62	0.73	0.80	0.48	0.75	0.59
MSCI: ESG	0.40	1.00	-0.02	0.34	0.18	0.38	0.67	0.57	0.45	0.68	0.33
Reprisk: ESG	-0.24	-0.02	1.00	-0.41	0.14	-0.33	0.03	-0.43	0.43	0.00	-0.14
SPGlobal: ESG	0.59	0.34	-0.41	1.00	0.06	0.69	0.67	0.84	0.46	0.69	0.60
TVL: ESG	0.12	0.18	0.14	0.06	1.00	0.10	0.46	0.18	0.45	0.45	0.56
Moody's: ESG	0.62	0.38	-0.33	0.69	0.10	1.00	0.73	0.85	0.46	0.74	0.50
AVG: ESG	0.73	0.67	0.03	0.67	0.46	0.73	1.00	0.84	0.84	0.99	0.73
PCA: ESG	0.80	0.57	-0.43	0.84	0.18	0.85	0.84	1.00	0.50	0.86	0.66
MAHA: ESG	0.48	0.45	0.43	0.46	0.45	0.46	0.84	0.50	1.00	0.80	0.62
AVG_{vote}: ESG	0.75	0.68	0.00	0.69	0.45	0.74	0.99	0.86	0.80	1.00	0.72
STV_{vote}: ESG	0.59	0.33	-0.14	0.60	0.56	0.50	0.73	0.66	0.62	0.72	1.00

Table 6.1: The cross-sectional rank correlation between ESG scores (individual and aggregate) averaged over time.

Region	FF5		FF3		CAPM	
	Mean	Std	Mean	Std	Mean	Std
USA	1.02	19.76	1.13	19.98	-0.55	22.93
Europe	6.55	21.46	6.14	21.60	3.86	23.35
Japan	4.27	18.64	3.29	19.13	0.14	20.72

Table 6.2: Alpha Statistics.
The cross-sectional mean and standard deviation of excess returns (alpha) of the sampled companies, computed using different factor models (FF5, FF3 and CAPM) averaged over time.

6.4 Performance of ESG Portfolios

In this section, we begin our study of the performance of ESG portfolios by first measuring the cross-sectional correlation between ESG scores and excess returns of individual stocks. We then summarize the empirical properties of both the raw returns and the returns in excess of Fama-French factor models for portfolios based on their individual and aggregated ESG scores. We then look at ESG portfolios constructed using Treynor-Black weights, and ESG portfolios combined with passive index portfolios.

6.4.1 Correlations between ESG ratings and Excess Returns

	USA			Europe			Japan		
ESG	FF5	FF3	CAPM	FF5	FF3	CAPM	FF5	FF3	CAPM
ISS: ESG	2.88	3.40	7.75	5.66	7.03	4.23	6.07	6.86	8.43
MSCI: ESG	4.46	4.77	6.70	3.74	5.13	5.38	5.41	5.09	6.23
Reprisk: ESG	2.75	2.36	2.71	-3.22	-1.37	6.11	2.40	1.88	4.45
SPGlobal: ESG	1.72	1.90	2.58	3.63	2.83	-3.66	6.41	5.51	6.01
TVL: ESG	2.17	2.92	3.43	2.42	4.43	5.62	3.63	3.47	2.24
Moody: ESG	2.81	2.82	4.38	3.22	3.12	-2.60	5.66	5.33	5.02
AVG: ESG	4.91	5.32	8.09	4.70	6.41	4.61	8.74	8.32	9.57
PCA: ESG	2.79	3.20	5.55	5.33	5.51	-0.36	6.73	6.57	6.93
MAHA: ESG	5.42	5.52	7.88	1.78	3.62	4.96	9.02	8.26	9.91
AVG_{vote}: ESG	4.90	5.40	7.67	5.11	6.69	4.62	8.85	8.66	9.89
STV_{vote}: ESG	3.59	4.18	5.45	4.31	4.86	0.86	8.25	7.85	7.74
OPT: ESG	6.06	6.14	9.32	6.56	7.99	5.32	9.13	8.47	10.37

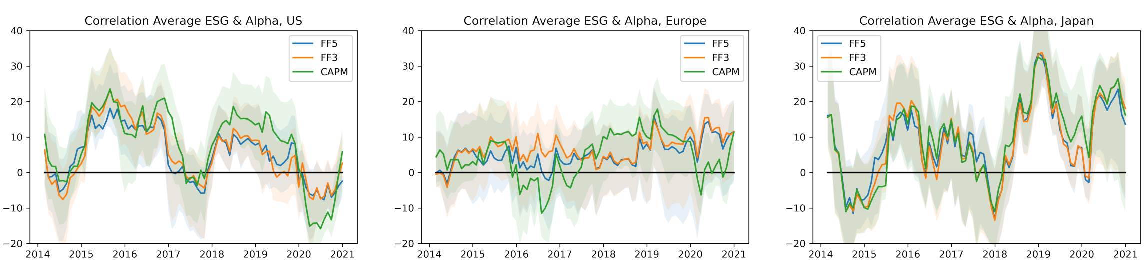
Table 6.3: Average cross-sectional correlation between ESG scores and alpha (excess returns).

Given the ESG scores of individual stocks and their excess returns, we compute the cross-sectional correlations, $corr_t = corr_i(ESG_{i,t}, \alpha_{i,t})$, on date t for different ESG scoring methods, where $\alpha_{i,t}$ is the one-year forward-looking alpha of stock i at date

t. The average values of $corr_t$ are presented in Table 6.3 for both individual and aggregated ESG scores. The last row shows that the maximum possible correlation (OPT: ESG) from linear combinations of individual ESG scores is in the range of 6 to 10% for all alphas and regions. Although the correlations for other aggregate ESG scores are inevitably lower, they are not much lower compared to the optimized ESG score. Different regions have different scores with the highest correlations. For example, in the U.S., MAHA: ESG and AVG: ESG achieve the highest correlations, while ISS: ESG has the highest correlations in Europe.

In Figure 6-1, we show the time series correlation values for AVG: ESG scores. We observe that correlations are typically in the range from 30 to 40% in the U.S. and Japan. However, there are some instances of negative correlation as well. A negative correlation implies that a high-low portfolio has negative excess returns. We find positive correlations in our sample on average because from 2014 to 2020, the ESG portfolio potentially had positive excess returns.

The different statistics described in Tables 6.2–6.3 can be used to obtain the average excess returns as described in Section 6.2.1. For example, the average alpha for the top decile/bottom decile ESG portfolio with a cross-sectional α deviation 20% and correlation value of 5% will be $2 * 0.20 * 5 * 1.64 = 3.28\%$. We discuss the excess returns of different ESG portfolios in Section 6.4.3 in more detail.



(a) (b) (c)

Figure 6-1: The correlation between average ESG score & stock performance (alpha) (y-axis) versus time (x-axis).
We find that the correlation between α and the average ESG score can be as high as 20% to 30% (in the U.S. and Japan, respectively); however, it clearly varies over time.

6.4.2 Returns and Factor Exposure

Following Section 6.2.2, we construct multiple ESG top-bottom portfolios (± 40, ± 25, and ± 10) and compute their properties, such as their returns, risk-adjusted returns (Sharpe ratios), excess returns (alphas) and exposure to different factors. In addition to portfolios based on the ESG scores from individual vendors, we also construct portfolios using the aggregate scores computed with the methods described in Section 6.2.3. We also include a portfolio which is simply the average portfolio constructed using individual ESG scores (represented by INDI-AVG: ESG).[14].

Table 6.4 presents the mean returns and annualized Sharpe ratios for ESG portfolios computed using individual vendor and aggregated ESG scores. The ESG portfolios achieve annualized returns as high as 6% in the U.S. and Japan, and 3% in Europe. Similarly, the highest Sharpe ratios are around 1 in the U.S. and Japan, and 0.7 in Europe. We find that portfolios based on ESG scores from individual vendors have nonuniform returns and Sharpe ratios. ESG scores from ISS, MSCI, and Reprisk have consistently positive returns across different regions, while portfolios based on scores from SPGlobal, TVL, and Moody's have negative returns in some regions. However, the portfolios using aggregated scores have positive returns, and are generally higher than returns of those based on constituent vendors.

Given the time series of portfolio returns, we follow Equation 6.1 and perform a time-series regression to compute the excess returns and fundamental factor exposure of these ESG portfolios. The response variable of the regression is the portfolio return, and the factors are chosen to be the Fama-French five factors.

We find that the ESG portfolios have varying exposures to multiple fundamental factors. We present the exposures in Table 6.5. We have a total of 33 portfolios (11 different ESG scores multiplied by 3 top-bottom portfolios). Table 6.5 presents the number of positive and negative significant betas (defined by a p-value $< 5\%$) out of 33 ESG portfolios. In the U.S. and Europe, ESG portfolios tend to have a

[14]This portfolio differs from the AVG: ESG portfolio because in AVG: ESG, we first compute the average ESG scores and then construct portfolios based on the average. However, in INDI-AVG: ESG, the first step is to construct portfolios based on individual ESG scores, and the second step is to average the returns of individual ESG portfolios.

	USA			Europe			Japan		
Portfolio	±40%	±25%	±10%	±40%	±25%	±10%	±40%	±25%	±10%
				Mean Returns					
ISS: ESG	4.89	4.57	5.51	2.37	1.66	1.10	4.68	5.76	6.98
MSCI: ESG	3.25	4.72	5.48	3.89	4.13	3.76	3.48	5.24	1.53
Reprisk: ESG	2.07	2.05	4.17	1.21	3.29	6.30	1.14	2.31	4.29
SPGlobal: ESG	-0.08	-0.48	1.46	-1.14	-0.70	2.02	2.62	3.63	7.78
TVL: ESG	2.39	2.49	-0.46	2.00	3.05	1.14	-0.52	-0.24	1.11
Moody: ESG	1.58	2.19	3.91	-0.93	-0.29	-1.89	2.35	1.86	6.39
INDI-AVG: ESG	2.35	2.59	3.34	1.23	1.85	2.18	2.29	3.09	4.68
AVG: ESG	3.32	6.48	6.39	2.60	3.21	4.30	5.26	5.38	4.83
PCA: ESG	2.26	3.69	2.94	0.28	0.68	1.54	3.04	3.81	4.20
MAHA: ESG	4.18	6.03	7.50	2.15	3.46	3.47	4.43	5.46	5.52
AVG_{vote}: ESG	3.18	5.56	6.85	2.29	3.49	4.46	4.99	6.13	5.29
STV_{vote}: ESG	2.59	2.07	1.30	-0.25	1.34	5.75	2.97	4.32	8.00
OPT: ESG	4.77	6.54	8.36	1.64	2.09	5.76	4.97	5.99	5.24
				Sharpe Ratio					
ISS: ESG	1.04	0.83	0.81	0.57	0.33	0.15	1.07	1.05	0.84
MSCI: ESG	1.07	1.23	0.94	1.00	0.86	0.53	0.93	1.05	0.19
Reprisk: ESG	0.53	0.45	0.67	0.27	0.57	0.73	0.25	0.39	0.51
SPGlobal: ESG	-0.03	-0.12	0.26	-0.28	-0.13	0.28	0.53	0.53	0.82
TVL: ESG	0.60	0.50	-0.07	0.58	0.75	0.21	-0.13	-0.05	0.17
Moody: ESG	0.45	0.45	0.55	-0.24	-0.06	-0.28	0.42	0.26	0.68
INDI-AVG: ESG	1.03	0.92	0.92	0.63	0.75	0.58	0.82	0.87	1.06
AVG: ESG	0.81	1.18	0.85	0.70	0.71	0.62	1.12	0.93	0.58
PCA: ESG	0.59	0.73	0.42	0.07	0.13	0.21	0.59	0.57	0.49
MAHA: ESG	0.99	1.09	1.04	0.61	0.79	0.51	1.17	1.08	0.73
AVG_{vote}: ESG	0.83	1.03	0.97	0.61	0.75	0.66	0.99	1.00	0.61
STV_{vote}: ESG	0.62	0.41	0.20	-0.07	0.31	0.96	0.67	0.70	1.03
OPT: ESG	1.14	1.25	1.22	0.43	0.43	0.79	1.15	1.07	0.67

Table 6.4: ESG Portfolio Performance

Mean returns and Sharpe ratios for the top $y\%$ minus bottom $y\%$ ESG portfolios (individual and aggregate scores) for the U.S., Europe and Japan, where $y = 40, 25, 10$.

negative exposure to the market and size factors, while having a positive exposure to the profitability factor. However, in Japan, ESG portfolios tend to have negative exposure to the profitability and the investment factors.

A varying exposure to different risk factors implies that the returns of the portfolios are not purely due to ESG risk premia. They may be due to exposure to different fundamental factors, as shown by the statistically significant coefficients in Table 6.5. Therefore, we next analyze the excess returns of the ESG portfolios.

		Market		Size		BM		Profitability		Investment	
		+ve	-ve	+ve	-ve	+ve	-ve	+ve	-ve	+ve	-ve
11*3 Portfolios	USA	0	8	3	22	4	13	4	0	1	2
	Europe	0	22	3	25	14	3	25	0	0	3
	Japan	5	0	0	2	4	9	0	17	0	18

Table 6.5: Factor Exposure
Number of significant positive (+ve) and negative (-ve) betas (defined by a p-value $< 5\%$) for all 33 top-bottom ESG portfolios (11 ESG scores by 3 portfolios).

6.4.3 Estimating Excess Returns

In this section, we evaluate the returns of ESG portfolios in excess of their Fama-French factors using two methods. The first method estimates the portfolio alphas based on Equation 6.6 in Section 6.2.1. Here, we use the theory first proposed by [186] to quantify the excess returns based on the correlation ρ between ESG scores and alphas of individual stocks. The second method estimates a forward-looking time-series regression of raw returns on Fama-French factors. We refer to the former as the model-implied alpha and the latter as the realized alpha. We compute the alphas of yearly top-bottom quantile portfolios using both methods, and show that the model-implied alphas match realized alphas very well.

We construct top-bottom quantile ESG portfolios in which stocks are sorted by individual ESG score. The stocks in each quantile portfolio are given equal weights. Therefore, for each score, we will have four portfolios at time t, denoted by Q1, Q2, Q3, and Q4. Figure 6-2 presents the scatter plot of model-implied and realized alphas. At time t, both the model-implied and realized alphas are computed for a one-year forward-looking window. The number of data points in the scatter plot is 6 (the number of ESG vendor scores) $\times$ 4 (the number of quantile portfolios) $\times$ 108 months in our sample = 2592. To robustly validate the proposed model, we compute the FF5, FF3 and CAPM alphas.

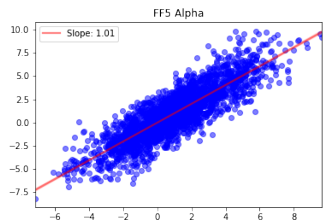
Figure 6-2 demonstrates that the realized and model-implied alphas of [186] match each other reasonably well, all data points falling around the line $y = x$. The slope of a simple linear regression of realized alphas against model-implied alphas is very close to 1 (p-value $< 1\%$) for all combinations of regions and factor models. Consequently, we only report the realized alphas using forward-looking time-series regressions henceforth.[15]

Next, we evaluate the annualized excess returns of the top-bottom ESG portfolios (± 40, ± 25 and ± 25), as described in Section 6.2.2. In Table 6.6, we present the excess returns (alphas) from the time series regression (Equation 6.1) using the Fama-French five-factor model and the Capital Asset Pricing Model. From Table 6.6, we find that the excess returns are positive and significant in the U.S. and Japan, but not in Europe. The CAPM alphas are generally higher than the FF5 alphas, implying that the Fama-French factors partially explain the positive returns beyond the market factor of the CAPM.

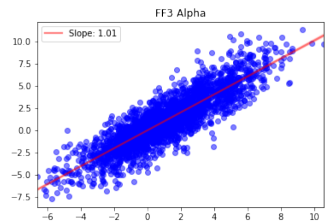
The excess returns of the OPT portfolios are comparable to the highest returns from those constructed using other aggregation methods or by individual scores. The excess returns from the ESG portfolios using OPT cannot be realized, however, but the other aggregate and individual ESG score alphas can. Hence, the highest realizable alpha is close to the maximum possible alpha that can be obtained using a linear combination of different ESG scores.

Among the individual scores, MSCI score portfolios consistently have the highest FF5 alphas across all regions, while ISS score portfolios have significant alphas in the U.S. and Japan. The aggregate ESG scores generally have higher alphas than most

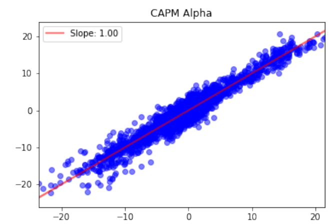
[15]The results using model-implied alphas are of course similar.



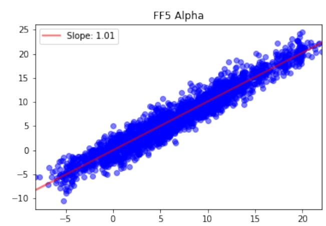
(a) **USA**, Alpha: **FF5**



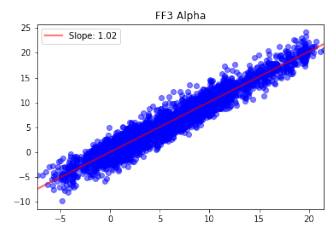
(b) **USA**, Alpha: **FF3**



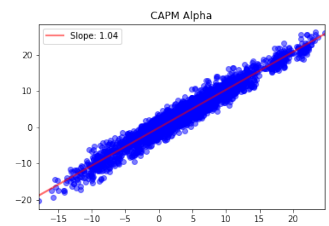
(c) **USA**, Alpha: **CAPM**



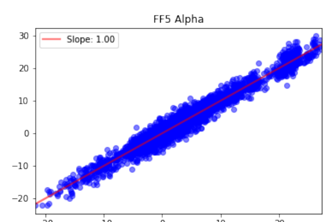
(d) **Europe**, Alpha: **FF5**



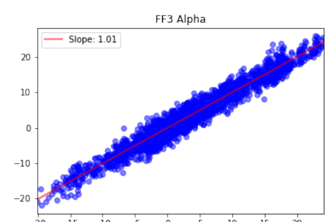
(e) **Europe**, **FF3**



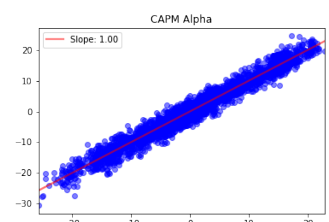
(f) **Europe**, **CAPM**



(g) **Japan**, **FF5**



(h) **Japan**, **FF3**



(i) **Japan**, **CAPM**

Figure 6-2: Realized alphas for portfolios created using ESG scores (y-axis) versus implied alphas (x-axis) for the same portfolios using the methods described in Section 6.2.1.

	USA			Europe			Japan		
Portfolio	±40%	±25%	±10%	±40%	±25%	±10%	±40%	±25%	±10%
FF5 Alpha									
ISS: ESG	3.30***	2.34*	3.01*	1.95	1.82	0.27	4.27**	5.96***	7.71***
MSCI: ESG	2.47**	3.80***	4.81**	2.90**	2.89**	1.49**	3.93**	5.84***	3.59***
Reprisk: ESG	0.95	0.86	2.75	-1.46	-0.40	0.80	0.09	1.27	2.88**
SPGlobal: ESG	0.72	0.31	1.92	0.43	1.22	3.26*	3.99***	5.11***	9.09***
TVL: ESG	2.37	1.79	-1.04	0.11	0.83	-1.28	-0.41	0.12	0.77
Moody: ESG	1.59*	2.73*	4.40**	0.58	1.29	-0.48	2.71*	1.94*	7.04**
INDI-AVG: ESG	1.90**	1.98**	2.64**	0.76	1.28	0.78	2.43**	3.37**	5.19***
AVG: ESG	2.69**	5.40***	4.50**	1.82*	2.32*	2.12*	5.58***	5.52**	5.92**
PCA: ESG	1.91*	3.23**	3.36*	1.70	1.93	2.51	3.92***	4.81**	5.17**
MAHA: ESG	3.44***	4.34**	6.09***	0.34	0.95	-0.49	4.14**	5.79***	7.25*
AVG_{vote}: ESG	2.58**	4.36***	5.44**	1.60	2.82*	2.85*	5.31***	6.33***	7.00**
STV_{vote}: ESG	2.96**	2.44**	0.88**	-0.49	1.39	6.06***	3.69**	5.36**	9.56***
OPT: ESG	3.87***	4.70***	5.73**	1.24	1.52	5.20**	5.32***	6.12***	6.47**
CAPM Alpha									
ISS: ESG	5.94***	5.39***	6.27***	2.46**	1.72**	1.31**	4.18**	5.22***	6.49**
MSCI: ESG	3.51***	5.00***	6.58***	4.11***	4.42**	4.21*	3.14**	4.79**	1.22**
Reprisk: ESG	1.86**	1.48**	2.89**	1.28	3.44*	6.57**	1.27	2.32	4.35***
SPGlobal: ESG	0.16	-0.26	2.26	-1.15	-0.75	1.96	2.09	2.87	6.56*
TVL: ESG	3.30**	3.02**	-0.30**	1.97	3.07*	1.14*	-1.20	-1.15	0.66
Moody's: ESG	2.14*	3.11**	5.10***	-0.92	-0.29	-1.96	1.32	0.41	4.82
INDI-AVG: ESG	2.82***	2.96**	3.80***	1.30**	1.94**	2.32*	1.79	2.40*	4.01**
AVG: ESG	4.23**	7.72***	7.41***	2.72**	3.34**	4.60**	4.58***	4.46**	3.71**
PCA: ESG	2.96**	4.76***	5.12**	0.32	0.75	1.73	2.35	2.88	3.17
MAHA: ESG	5.16***	6.89***	8.60***	2.29*	3.66**	3.80**	3.98**	4.76**	4.75**
AVG_{vote}: ESG	3.91***	6.55***	7.25***	2.41*	3.63**	4.76**	4.23**	5.03**	3.93**
STV_{vote}: ESG	3.75**	3.11**	1.34**	-0.21	1.37	5.80***	2.46	3.56	7.07**
OPT: ESG	5.76***	7.49***	8.78***	1.69	2.17	5.99***	4.58***	5.17**	4.11**

Table 6.6: Excess Returns for ESG Portfolios
FF5 and CAPM alphas for top-bottom ESG portfolios (±40%, ±25%, ±10%). The alphas are computed using time series regression. The standard errors are computed using heteroskedastic and autocorrelation consistent standard error estimators, with statistical significance highlighted at the 1% (***), 5% (**), and 10% (*) levels.

individual scores. These portfolios behave differently across regions and quantiles. For example, the STV_{vote} voting aggregate score generally has high ± 10 portfolio alphas in Europe and Japan. Broadly speaking, the alphas of different aggregation methods are similar to each other, due to high correlation between their scores.

From Table 6.6, we find there are significant excess returns of 4.8%, 2.9% and 9% using individual ESG scores in the U.S., Europe, and Japan, respectively, while there are excess returns of 6%, 6% and 9.6% using aggregate ESG scores. The alpha of portfolios based on Reprisk, TVL, Moody's scores are not statistically significant (possibly due to the noise in the scores). However, the alphas for the aggregate ESG portfolios are in general significant, likely due to the stronger signal from the aggregation methods.

6.4.4 Treynor-Black Portfolios

Along with equal-weighted portfolios, we also construct ESG portfolios using Treynor-Black weights, as given by Equation (6.7) in Section 6.2.2. In Treynor-Black portfolios, the weights are inversely proportional to the rank of the ESG score of each firm. In Table 6.7, we include the excess return (alpha) obtained from the time series regression using the Fama-French five-factor model and the Capital Asset Pricing Model.

Comparing the results in Tables 6.7 and 6.6, we do not observe a large difference between the alphas of the ± 25 and ± 10 portfolios, but we do find differences in alphas for the ± 40 portfolios. The effect of unequal weighting becomes more prominent when more firms are included in the portfolio and when firms at extreme percentiles on the long or short sides are weighted differently. Our other observations about excess returns described in Section 6.4.3 remain consistent.

6.4.5 Combining ESG and Passive Portfolios

Once the relative weights of securities within an ESG portfolio are determined, one can combine that portfolio with any other portfolio. For example, we can also add

	USA			Europe			Japan		
Portfolio	±40%	±25%	±10%	±40%	±25%	±10%	±40%	±25%	±10%
				FF5 Alpha					
ISS: ESG	2.62***	2.20*	2.37*	2.42	2.43	1.43	5.40**	5.94***	7.03***
MSCI: ESG	3.31**	3.92***	4.49**	2.75**	2.88**	1.74**	4.66**	5.51***	3.62***
Reprisk: ESG	1.52	1.52	3.31	-0.64	0.02	1.47	1.12	1.73	2.44**
SPGlobal: ESG	0.93	0.62	1.66	2.00	2.48	3.80*	5.30***	5.95***	8.61***
TVL: ESG	1.19	0.66	-1.59	0.21	0.54	-1.28	-0.17	0.11	0.93
Moody's: ESG	2.31*	2.89*	3.94**	1.13	1.51	0.58	2.90*	2.80*	6.91**
INDI-AVG: ESG	1.98**	1.97**	2.37**	1.32	1.66	1.32	3.20**	3.67**	4.95***
AVG: ESG	4.09**	5.42***	4.64**	1.94*	2.16*	1.26*	5.87***	5.96**	6.37**
PCA: ESG	2.40*	2.88**	2.80*	2.46	2.81	3.28	5.02***	5.55**	5.90**
MAHA: ESG	4.01***	4.31**	5.17***	0.87	1.08	-0.32	5.49**	6.37***	7.62*
AVG_{vote}: ESG	3.81**	4.61***	5.12**	2.19	2.71*	2.35*	6.04***	6.63***	7.44**
STV_{vote}: ESG	1.85**	1.55**	-0.14**	1.95	3.02	6.73***	5.51**	6.52**	9.94***
OPT: ESG	4.68***	5.18***	6.19**	3.05	3.47	6.33**	6.33***	6.81***	7.08**
				CAPM Alpha					
ISS: ESG	5.60***	5.37***	5.63***	3.05**	2.85**	2.93**	4.96**	5.24***	6.07**
MSCI: ESG	4.68***	5.37***	6.49***	4.34***	4.64**	4.34*	3.41**	4.10**	1.11**
Reprisk: ESG	2.04**	1.88**	3.33**	3.15	4.40*	7.34**	2.24	2.77	3.68***
SPGlobal: ESG	0.45	0.24	1.92	0.39	0.73	2.63	3.22	3.67	6.10*
TVL: ESG	2.12**	1.65**	-0.98**	2.21	2.78*	1.12*	-1.12	-1.05	0.59
Moody's: ESG	2.82*	3.33**	4.47***	-0.09	0.25	-0.38	1.33	1.15	4.78
INDI-AVG: ESG	2.95***	2.98**	3.48***	2.19**	2.63**	3.03*	2.33	2.64*	3.73**
AVG: ESG	6.11**	7.79***	7.32***	3.43**	3.80**	4.10**	4.40***	4.37**	3.79**
PCA: ESG	3.67**	4.28***	4.32**	1.54	2.02	2.89	3.30	3.68	4.00
MAHA: ESG	6.10***	6.80***	7.54***	3.64*	4.27**	4.20**	4.49**	4.84**	4.82**
AVG_{vote}: ESG	5.59***	6.74***	6.85***	3.44*	4.04**	4.67**	4.44**	4.81**	4.15**
STV_{vote}: ESG	2.57**	2.19**	0.36**	2.12	3.01	6.49***	3.76	4.44	7.37**
OPT: ESG	7.09***	8.07***	9.20***	3.69	4.20	7.22***	5.04***	5.32**	4.46**

Table 6.7: Excess Returns for Treynor Black ESG Portfolios
FF5 and CAPM alphas for Treynor-Black top-bottom ESG portfolios (±40%, ±25%, ±10%). The alphas are computed using time series regression. The standard errors are computed using heteroskedastic and autocorrelation consistent standard error estimators, with statistical significance highlighted at the 1% (***), 5% (**), and 10% (*) levels.

the ESG portfolio to a suite of portfolios that mimic more traditional asset pricing factors, such as value, size, or momentum.

Perhaps the most natural application is to combine the ESG portfolio with a passive index fund such as the market portfolio. In this section, we combine the active Treynor-Black ESG portfolios with market portfolios.[16] The weight of the market portfolio is fixed to be 0.5.[17] The market portfolios we use for different regions are the MSCI USA, the MSCI Europe, and the MSCI Japan indices.

Figure 6-3 presents the expected return and volatility of several combined ESG portfolios. In each region, we include the top-bottom portfolios (± 40, ± 25, and ± 10) with the highest Sharpe ratio, based on single or aggregated ESG scores. The Sharpe ratios of the combined portfolios are higher than those of market portfolios, due to the signal in the ESG scores. These improved Sharpe ratios are not accessible to traditional mean-variance optimized portfolios which stay below the capital market line. This forms a "super-efficient frontier" compared to the capital market line associated with the passive portfolio, assuming that the alphas from the ESG portfolios are mispricings. Under the alternate interpretation that ESG scores capture an omitted pricing factor, the "super-efficiency" of the new frontier may be viewed as the result of additional risk premia not accessible to investors except for ESG portfolio managers.

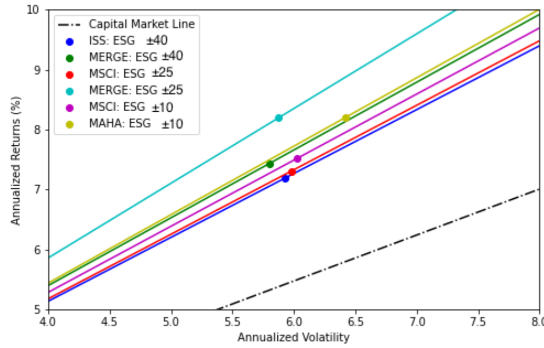
In particular, the combined portfolios achieve annual Sharpe ratios reach as high as 1.25, 0.53 and 0.72 in the U.S., Europe, and Japan, respectively. As a comparison, the Sharpe ratios of the market portfolios are only 0.75, 0.05, and 0.37. Across all combined portfolios across ESG scores, the Sharpe ratios are always positive in the U.S. and Japan, while they are negative for portfolios based on Moody's scores in Europe.

[16] Table E.3 in the supplementary material gives the Sharpe ratio for the portfolios that are built by combining equal-weighted ESG portfolios and the passive market index.

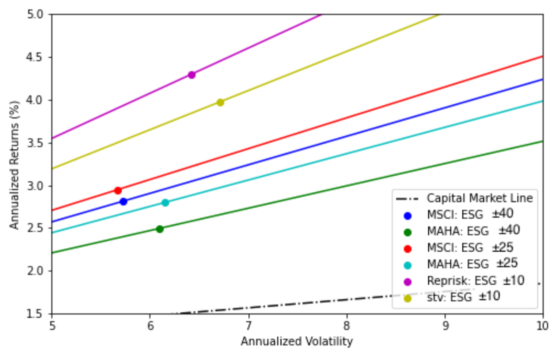
[17] More generally, weights can also be determined by other methods. For example, [186] show that the optimal weights to maximize the Sharpe ratio ω can computed by using an ESG portfolio's excess return and idiosyncratic volatility:

$$\omega_A = \left(\frac{\alpha_A}{\sigma(\epsilon_A)^2}\right) \Big/ \left(\frac{\mathbb{E}[R_m] - R_f}{\sigma_m^2}\right) \tag{6.12}$$

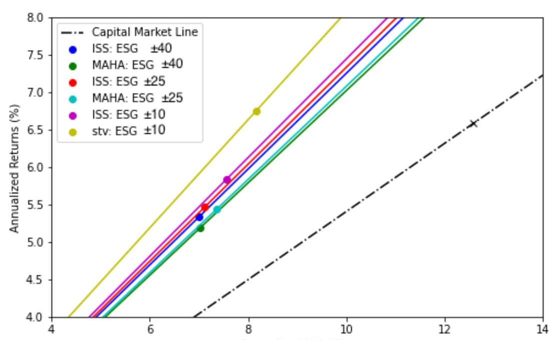
where $\mathbb{E}[R_m]$ and σ_m^2 are the expected return and variance of the passive portfolio, respectively.



(a) USA



(b) Europe



(c) Japan

Figure 6-3: Annualized returns (y-axis) versus volatility (x-axis).
We present the return versus volatility for portfolios that combine a Treynor-Black ESG portfolio with the market portfolio. The capital market line is represented using the black dotted line. The combined portfolios lie above the capital market line. The entries on the legend are in the format: “ESG Dataset portfolio name"

6.5 ESG as Univariate Impact Factors

We now turn to analyzing portfolios constructed using univariate Environment (E), Social (S), and Governance (G) scores. Like our analysis of portfolios based on full ESG scores, we construct top-bottom sorted portfolios using the E, S and G scores individually, and compute their excess returns.

6.5.1 Environment Portfolios

We include the excess returns for Environment portfolios in Table 6.8. For Environment scores, we observe high and statistically significant alphas (up to 10.5%) for multiple individual vendor scores and aggregate scores for Japan. For the U.S., individual vendor scores do not achieve significant alpha, while aggregate scores generate significant positive alpha, with excess returns of up to 4%. For Europe, we do not observe significant alphas. However, CAPM alphas are positive, significant and higher for the U.S. compared to other regions.

6.5.2 Social Portfolios

We find there are similar patterns for Social score portfolios (see Table 6.9). However, these portfolios have negative significant excess returns (CAPM alpha) in Europe. Fama-French five-factor annualized alphas for Social score portfolios are positive and significant in the U.S. (up to 4.3%) and Japan (up to 7.48%).

6.5.3 Governance Portfolios

For Governance score portfolios (see Table 6.10, the aggregation does not produce significant Fama-French five-factor alphas in the U.S. or Europe compared to individual vendor scores; however, we find there are Fama-French five-factor excess returns of up to 11.75% in Japan.

	USA			Europe			Japan		
Portfolio	±40%	±25%	±10%	±40%	±25%	±10%	±40%	±25%	±10%
				FF5 Alpha					
ISS: E	2.46*	0.76*	-0.17*	1.79	1.13	-0.04	3.21	4.19*	8.21***
MSCI: E	3.12*	4.24*	8.72**	0.73	0.24	0.16	3.68***	4.55***	6.54***
Reprisk: E	0.55	2.22	0.45	-1.70	-2.65*	-0.90*	3.06*	2.53*	1.41*
SPGlobal: E	1.60	1.66	2.20	0.01	1.36	4.02**	4.34***	5.82***	10.45***
Moody's: E	0.74	1.39	-2.35	2.05**	1.26**	-0.58**	2.11	1.83	4.42
AVG: E	2.98**	3.95**	4.49**	0.63	0.42	0.84	3.45*	7.30***	9.99***
PCA: E	2.50**	1.78**	-2.10**	0.78	1.37	0.48	3.71***	4.31***	7.54***
MAHA: E	1.67	3.30*	5.83*	1.37	1.16	0.50	4.69**	7.37***	9.65***
AVG_{vote}: E	2.29*	3.34*	4.16*	1.09	1.61	1.63	3.40*	6.94***	8.88***
STV_{vote}: E	1.95	1.88	2.26	0.06	0.33	3.53*	5.15***	6.74***	10.65***
				CAPM Alpha					
ISS: E	4.96***	3.37***	2.93***	2.44**	2.15**	2.30**	2.97	3.99*	7.24**
MSCI: E	5.63**	7.72**	13.68***	0.75	0.51	0.97	2.55	3.39*	5.30*
Reprisk: E	1.66	3.17*	1.92*	0.63	0.32	4.27	4.74***	4.19**	4.09**
SPGlobal: E	2.69*	2.50*	3.85**	-1.26	-0.30	2.67*	1.90	3.34*	7.87**
Moody's: E	1.94**	3.12***	-1.31***	1.05	-0.04	-1.03	0.57	0.15	3.93*
AVG: E	5.25***	7.50***	10.27***	1.39	0.78	2.13	2.62*	5.86***	8.42**
PCA: E	4.22***	3.82**	0.14**	-0.27	0.14	-0.55	1.88	2.33	6.03**
MAHA: E	3.67***	6.32***	10.96***	2.62	3.26	3.33	4.50***	6.86***	8.95***
AVG_{vote}: E	4.47***	6.61***	10.05***	1.27	1.63	2.73	2.17	5.40**	7.66**
STV_{vote}: E	3.19*	3.13*	5.51***	-1.03	-0.88	2.70	3.26**	4.61**	7.66**

Table 6.8: Excess Returns for E Portfolios

FF5 and CAPM alphas for top-bottom **E** portfolios (±40%, ±25%, ±10%). The alphas are computed using time series regression. The stars next to the numbers represents significance levels. ***: p-val< 0.01, **: p-val<0.05, *: p-val < 0.10. The standard errors were computed using heteroskedastic autocorrelation consistent standard error estimators.

	USA			Europe			Japan		
Portfolio	±40%	±25%	±10%	±40%	±25%	±10%	±40%	±25%	±10%
				FF5 Alpha					
MSCI: S	0.86	0.37	3.04**	1.70	1.95	-0.62	2.82**	2.66**	2.22**
Reprisk: S	1.76	1.84	0.05	-0.72	-0.14	1.20	0.43	-0.60	-1.28
SPGlobal: S	1.42	1.58	-3.40	-0.78	0.15	-1.61	3.87**	5.12***	7.61***
Moody's: S	2.56**	3.28**	1.67**	-0.67	-0.80	1.60	3.44**	0.94**	3.13**
AVG: S	2.69***	3.64***	3.83*	0.44	0.11	-2.32	2.64**	3.81*	7.10**
PCA: S	1.81	1.51	-0.30	-0.88	0.20	0.87	2.79**	3.89***	5.10**
MAHA: S	3.93***	4.28***	2.54*	0.64	1.13	-1.95	3.00**	3.01*	6.19*
AVG_{vote}: S	2.29***	2.51**	4.30***	0.98	-0.48	-2.32	3.45**	4.97**	5.81*
STV_{vote}: S	1.31	1.49	-1.37	-0.43	0.31	-1.36	3.64**	5.68***	7.48***
				CAPM Alpha					
MSCI: S	0.90	-0.40	1.49	1.11	1.58	-0.80	1.19	0.57	0.05
Reprisk: S	2.35**	2.13**	-1.29**	1.71	2.98*	5.41*	1.96	1.53	1.21
SPGlobal: S	1.46	1.43	-3.20	-2.34***	-2.01*	-3.23*	1.43	3.15	5.13
Moody's: S	3.43**	4.23**	2.71**	-2.59***	-2.62**	-1.14**	1.18	-1.03	1.51
AVG: S	3.49***	4.50***	3.98*	-0.40	-0.82	-1.56	1.23	2.11	4.73
PCA: S	2.49*	1.67*	-0.06*	-3.03***	-2.50***	-2.84***	0.34	1.02	1.95
MAHA: S	5.19***	5.97***	3.29*	0.78	1.77	-0.14	1.95	2.62	4.60
AVG_{vote}: S	2.88***	2.71**	4.28**	-0.37	-1.62	-2.25	1.79	2.57	3.20
STV_{vote}: S	1.37	1.67	-0.47	-1.93***	-1.76***	-2.72***	1.31	3.61	6.00

Table 6.9: Excess Returns for S Portfolios.

FF5 and CAPM alphas for top-bottom **S** portfolios (±40%, ±25%, ±10%). The alphas are computed using time series regression. The stars next to the numbers represents significance levels. ***: p-val< 0.01, **: p-val<0.05, *: p-val < 0.10. The standard errors were computed using heteroskedastic autocorrelation consistent standard error estimators.

	USA			Europe			Japan		
Portfolio	±40%	±25%	±10%	±40%	±25%	±10%	±40%	±25%	±10%
				FF5 Alpha					
ISS: G	1.89**	2.32*	3.98*	-0.37	0.31	-0.19	4.18***	4.12***	3.91**
MSCI: G	2.43**	3.66***	4.01***	1.08	0.17	1.90	1.99	2.99**	5.23***
Reprisk: G	0.34	2.32**	1.79*	0.51	0.71	4.60*	-0.14	1.34	4.89***
SPGlobal: G	-0.25	0.42	2.72	-0.13	1.00	3.98**	3.45**	5.12***	11.75***
Moody's: G	0.29	1.21	1.30	0.83	0.45	0.83	1.71	2.88	3.09
AVG: G	2.25**	3.44**	5.84**	0.75	0.81	0.31	4.44***	5.01***	6.68*
PCA: G	2.05	1.39	0.86	0.26	1.28	2.79	0.18	-0.44	-3.86
MAHA: G	1.57	1.91	4.27*	0.83	0.54	-1.33	3.33*	5.44***	3.45***
AVG_{vote}: G	1.74*	2.43*	3.72*	0.79	0.85	0.31	4.50***	4.85**	7.42***
STV_{vote}: G	-0.28	0.40	2.58	0.18	1.36	3.98**	3.46**	5.05***	9.79***
				CAPM Alpha					
ISS: G	2.32**	2.65**	4.02**	1.15	1.79	1.20	3.60***	3.24***	2.40***
MSCI: G	1.51	2.51**	1.91**	3.09***	3.34**	7.46**	1.57	2.78	5.31***
Reprisk: G	0.42	2.22**	1.19**	2.61	3.36*	8.61***	0.85	2.89	6.78***
SPGlobal: G	-2.26	-0.82	3.30	-1.86*	-1.69**	0.29**	1.15	1.80	7.56***
Moody's: G	-0.08	0.60	-0.43	0.67	-0.13	1.36	-0.29	0.61	-1.06
AVG: G	1.42	2.19	4.03	2.64***	3.51*	2.45*	3.39**	3.36*	5.58*
PCA: G	1.49	0.58	0.19	-0.27	1.03	2.00***	-0.33	-0.48	-3.98*
MAHA: G	1.01	0.69	1.95	2.71**	3.33**	1.48**	2.77*	4.54**	2.62**
AVG_{vote}: G	0.60	1.20	2.06	2.65***	3.38**	1.74**	3.25**	2.98*	4.70*
STV_{vote}: G	-2.22	-0.55	2.50	-1.36	-1.22	0.87	1.29	2.02	5.82**

Table 6.10: Excess Returns for G Portfolios.

FF5 and CAPM alphas for top-bottom **G** portfolios (±40%, ±25%, ±10%). The alphas are computed using time series regression. The stars next to the numbers represents significance levels. ***: p-val< 0.01, **: p-val<0.05, *: p-val < 0.10. The standard errors were computed using heteroskedastic autocorrelation consistent standard error estimators.

6.6 Conclusion

Using the quantitative framework proposed by [186], we quantify the excess returns of arbitrary ESG portfolios via the cross-sectional standard deviation of the stock's excess returns and the correlation between the excess return and ESG factors (both combined and individual E, S, and G scores) obtained from six leading ESG score providers for firms in the U.S., Europe and Japan from 2014 to 2020. Few studies have analyzed such a comprehensive dataset and as systematically. We also propose a number of methods to aggregate ESG scores across vendors to produce the best signal within the data, simultaneously addressing measurement errors and yielding a single measure of ESG that can potentially be used for portfolio management.

Empirically, we find significant ESG excess returns in the U.S. and Japan. We also find positive and higher than market risk-adjusted returns. We construct an aggregate ESG measure based on a linear combination of ESG scores that is optimized to maximize the correlation with excess returns. The ESG portfolio properties of the optimized ESG score are comparable to the aggregate scores, implying that our methods of aggregation were successful in amplifying the signal. We evaluate the properties of ESG portfolios by investigating their exposure to various risk factors, constructing optimal Treynor-Black-weighted portfolios, and combining them optimally with passive index portfolios, which yields "super-efficient" frontiers that all investors should be interested in accessing.

One practical implication from our results is that aggregation methods help to reduce the noise and amplify the signal contained in E, S, and G metrics to yield better estimates of ESG portfolio properties, even though individual ESG ratings are noisy and the portfolios constructed using ESG scores from any single vendor may be quite noisy. This aggregation method can be selected at the preference of the portfolio manager, since different methods will weight the noise and signal from rating agency scores in different ways. However, since the true noise and signal component remains unknown, it is hard to establish the superiority of any particular aggregation method and we leave this important topic for future research.

Chapter 7

Finance for Impact: From Pills to Power: Unveiling the Impact of BioPharmaceutical Companies

7.1 Introduction

Environmental, Social, and Governance (ESG) scores are measures that assess a company's performance in the respective areas. The Environmental component assesses the company's impact on the environment by measuring its carbon footprint, managing pollution, conservation of resources, etc. The Social component measures performance on social issues such as labor practices, human rights, diversity, and inclusion, amongst others. Similarly, the Governance score measures the company's policies and evaluates how well the company complies with regulatory and ethical requirements. ESG scores are necessary because they provide a framework for stakeholders and investors to evaluate sustainability and social responsibility practices.

In recent years, ESG scores have gained significant interest from investors. This can be gauged from the fact that the number of investor signatories to the United Nations Principles of Responsible Investment has increased significantly, and assets under management have grown from 40 trillion USD in 2014 to 103 trillion USD in

2020.[1]

The sources of ESG evaluations/ratings for investors are rating agencies like MSCI Inc, S&P Global, ISS, Moody's ESG Solutions, Reprisk, Refinitiv, and TruValue Labs. They have developed methodologies to extract ESG information from multiple sources and generate ratings for companies. These rating agencies compute a variety of factors; for example, MSCI Inc considers aspects like toxic waste, water stress, carbon footprint, electronic waste, opportunities in clean tech, and biodiversity & land use when computing the environmental score.

Pharmaceutical companies specialize in the research and development of drugs and medications for various diseases. Their products contribute significantly to decreasing the burden of disease worldwide. For example, an article from Medscape claims that COVID-19 vaccines have saved more than 3 million US lives since 2020.[2] To summarize, the treatment of disease is an important social cause to which pharmaceutical companies contribute significantly.

In the computation of ESG scores for pharmaceutical companies, some rating agencies account for factors specific to the pharmaceutical industry. For example, the MSCI ESG score includes product safety & quality, access to health care, and chemical safety when computing the social score, while S&P Global and Sustainalytics consider aspects like drug affordability when computing the ESG score. However, none of the rating agencies currently take into account the impact of the drugs/treatments produced by pharmaceutical companies on patients' lives.

In this chapter, we develop a framework that quantifies the impact of therapeutics developed by pharmaceutical companies in terms of reducing the burden of diseases. We collect data from multiple sources and employ a quantitative tool to estimate the burden that will be reduced by the developed therapeutics. In this work, we quantify the reduced burden by the decrease in Disability-Adjusted Life Years (DALYs) due to a therapeutic.

Quantifying the impact of therapeutics will allow investors to accurately gauge the

[1]https://www.unpri.org/annual-report-2020/how-we-work/building-our-effectiveness/enhance-our-global-footprint

[2]https://www.medscape.com/viewarticle/985529

social impact of pharmaceutical companies. This will also lead to an inflow of ESG funds, enabling access to capital for financing new research and the development of more therapeutics. It will also allow the identification of pharmaceutical companies with high impact scores, leading to enhanced reputation and increased trust from stakeholders, including investors, customers, and regulators.

From the impact score measures we developed, we found that major pharmaceutical companies like Pfizer, Sanofi, Novartis, GlaxoSmithKline, and Johnson & Johnson had the highest impact scores (in decreasing order), which is intuitive given the large number of launched drugs. We also constructed market capitalization normalized scores and found that smaller companies like NeuroHitech, AVAX Technologies had high scores alongside larger companies like GlaxoSmithKline. We found that cardiovascular diseases, cancer, diabetes, and musculoskeletal disorders contribute the most to the impact measures. We also found that our impact scores are not correlated with E, S, G, and ESG scores released by Refinitiv, which implies that existing ESG scores from Refinitiv are not capturing the social impact of the biopharmaceutical companies.

The study of impact scores and ESG is a widely examined topic. However, the impact scores pertaining to pharmaceutical companies have not been explored in significant detail. The Global Health Impact project quantified the impact of drugs for major diseases like malaria, tuberculosis, and HIV/AIDS based on historical sales data, but it doesn't extend this analysis to all diseases. Existing ESG vendors consider drug affordability but don't compute the impact score of the drugs. Similarly, some organizations like the Institute for Clinical and Economic Review (ICER) and IQVIA perform cost-effectiveness analyses for a few drugs. In this work, we estimate the impact scores of therapeutics, a factor not explored in existing literature (to the best of our knowledge).

In Section 7.2, we discuss the datasets and impact estimation methodology. In Section 7.3, we study our impact scores and analyze them. We conclude in Section 7.4 with a discussion about the shortcomings and potential future extensions."

7.2 Methodology

7.2.1 Data

Multiple datasets were utilized to quantify the impact of biopharmaceutical companies. These datasets are enumerated as follows:

We obtained the drug dataset from Pharmaprojects (Citeline-Informa)[3]. This is a widely used dataset that contains information about the drug, its development status, sponsor, drug-indication, among other information.

Summary	Number of Unique
Drugs	6246
Country	157
Companies	1398
Indications	1187

Table 7.1: Informa Data Summary for Launched Drugs.

The burden of disease can be quantified using multiple measures, and one of the most widely used measures is Disability-Adjusted Life Years (DALYs). This measure represents the number of years lost due to ill health, disability, or early death. The dataset on the burden of disease is obtained from the Global Burden of Disease (GBD) study [219]. This study is one of the most comprehensive systematic evaluations of the global burden of disease and reports multiple measures such as mortality, years of life lost (YLLs), years of life lived with disability (YLDs), and DALYs for diseases grouped in a hierarchy. Diseases are classified based on the International Classification of Diseases, Tenth Revision (ICD-10) codes, which is the system used by physicians to classify and group diseases.

The market capitalization of the companies was obtained from the Center for Research in Security Prices (CRSP) dataset. Effective coverage was obtained from the study by [191]."

[3]https://pharmaintelligence.informa.com/products-and-services/clinical-planning/pharmaprojects

Summary	Number of Unique
Locations	204
Causes	364

Table 7.2: GBD Data Summary.

Pre-processing

The datasets were obtained from different sources. Given these datasets, one of the challenges was to match data entries across them, as there was no unique identifier common across all datasets. The necessary mappings included: companies between the CRSP and drug dataset; diseases between the drug and GBD datasets; and countries between the GBD, drug dataset, and effective coverage.

We used fuzzy string matching methods to map the companies across datasets. Along with fuzzy clustering, we manually verified the matches to ensure accurate mapping. The disease names between datasets were mapped using the ICD-10 codes. The countries were matched using string matching and manual verification of failed matches.

7.2.2 Pipeline

The impact of a drug on a disease can be measured by counting the DALYs averted, which is computed by the following equation:

$$
\begin{aligned}
Effect_{D,I,C} &= DALY_{I,C} * (\frac{1}{1 - efficacy_{D,I} * Coverage_C} - 1) \\
Impact_{D,I,C} &= Effect_{D,I,C} * MarketShare_{D,I,C} \\
Impact_{Company} &= \sum_{D,I,C} Impact_{D,I,C}
\end{aligned}
$$

where D represents the Drug, I represents the indication, and C represents the company. DALY represents the burden of the indication treated by the drug, efficacy is the ability of the drug to reduce the burden for the specific indication, market share represents the percentage of market captured by the drug for the specific indication, and coverage is the availability of intervention/treatment for the particular indication in the country.

The impact score for the company is computed as the sum of the impact over all approved drugs for a specific indication in the particular country.

Obtaining DALY for the indication is straightforward from the GBD dataset. Coverage is obtained from the UHC coverage indicator. The information about drugs and their indications is obtained from the GBD dataset. However, computing the market share of a drug is not trivial. The precise market share can be computed based on the number of doses sold, however, the number of doses needed for treatment varies among drugs. Hence, computing market share is a complex problem. We approximate the market share using two models:

- **Equal Coverage**: In this model, we assume equal market share for all drugs for a country-indication. For example, if there are five launched drugs for an indication in a country, we assume a market share of 20% for each of them.

- **Age Based**: In this model, we assume that market share depends on the age of the drugs. We model the market share using the Bass model which states the following:

$$f(t) = p + (q - p)F(t) - qF(t)^2$$

  where, F(t) is the cumulative number of adopters as a fraction of market potential and f(t) is the number of new adopters as a fraction of market potential. The Bass model is inspired by [233] as pharmaceutical companies commonly use it for predicting sales of new drugs. p and q are the parameters obtained from [233]. In Figure 7-1, we present the values of market share obtained from solving the Bass model differential equation. The pattern suggests that market share increases from launch to the first 5 years and starts decreasing after 5-6 years.

Once we have the necessary estimates[4], we compute the effect of drugs, the impact

[4] There is no publicly available dataset to obtain the efficacy scores. In our study, we used an

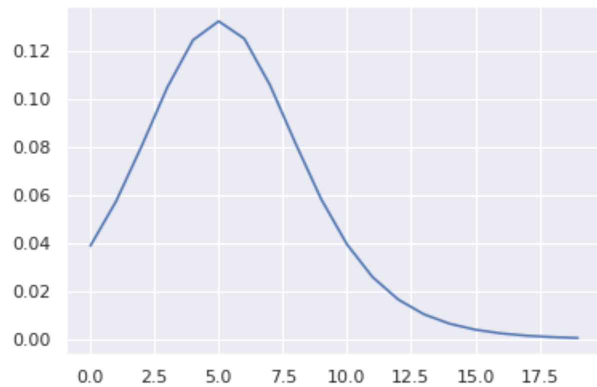


Figure 7-1: Bass Model. x-axis: year from launch, y-axis: f(t), number of new adopters.

of drugs, and aggregate the impact scores of drugs to compute the scores for the companies.

7.3 Analysis of Scores

After computing the impact scores, we analyze their properties in this section. Given that we have two models for market share computation, we also have two versions of impact scores: those based on equal and age-based market share.

In Table 7.3, we include the companies with the highest impact scores based on both models. We observe that big pharmaceutical companies are on the list, which is intuitive given that they have multiple drugs on the market treating multiple diseases, which contributes to the impact scores. However, we find that there is no significant difference in impact scores obtained from both models. To verify this, we compute the correlation between impact scores from both models and find a very high correlation: 98% between raw scores from both models and 95% between log-transformed scores from both models. Hence, the choice of the market share model doesn't significantly impact the scores. For further analysis, we will use the equal market share model for

efficacy score of 75% for all drugs. The estimates can be made more accurate by using the efficacy data from FDA labels. However, this will only be available for drugs/treatments approved in the U.S.A.

the computation of the impact score."

Company	Impact Score (Equal) *10^7	Impact Score (Age) *10^7
Pfizer	16.4	13.2
Sanofi	13.0	10.2
Novartis	11.4	10.6
GlaxoSmithKline	10.1	8.9
Johnson & Johnson	7.5	8.4
Astrazeneca	7.0	6.9

Table 7.3: Companies with High Impact Scores

In Figure 7-2, we include the histograms of the raw and log-transformed impact scores for the companies. The log-transformed impact score distribution resembles a bell curve with a lower range.

Given that large companies have multiple drugs on the market, we also compute the market capitalization normalized scores, which measure the impact score (DALYs reduced) per unit of the company's market capitalization. In Table 7.4, we present the impact scores normalized by the companies' market capitalization, i.e., $\frac{log(ImpactScore)}{log(MarketCapitalization)}$. We computed the normalized score for 238 companies with available market capitalization values. We find that some companies, like NeuroHitech and AVAX Technologies, have low absolute impact scores but high market cap normalized impact scores because they have only one approved drug on the market and their market capitalization values are relatively low. NeuroHitech has an approved drug for people with Alzheimer's disease or other types of dementia in China, and AVAX Technologies has one approved drug for Neoplasms in Switzerland. However, GlaxoSmithKline has high raw and normalized impact scores. It has a total of 303 launched drugs on the market.

The biopharmaceutical companies are broadly divided into two categories (based on the Global Industrial Classification System): Biotechnology and Pharmaceuticals. We computed the average impact scores (log-transformed version) for Biotechnology and Pharmaceutical companies and found that the average score for biotech companies (11.8$\pm$3.1) was smaller than for pharmaceutical companies (13.8$\pm$3.2). However, the difference was not statistically significant.

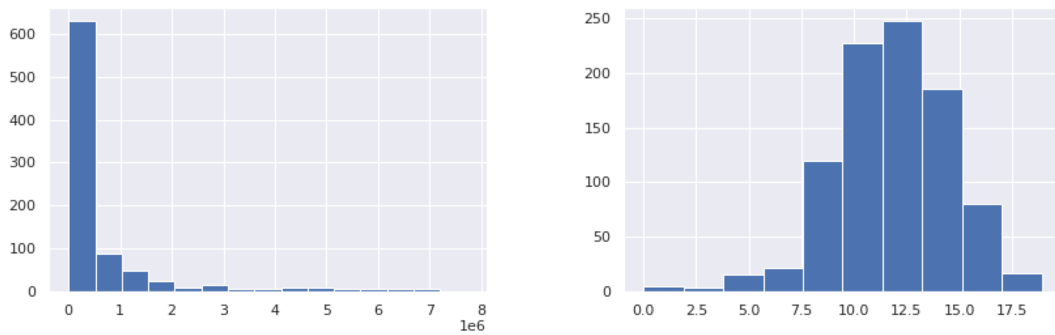


Figure 7-2: Histograms of Impact Scores of the Companies. Left: Raw Score, Right: Log Transformed Scores. x-axis: Impact Scores, y-axis: Number of Companies.

Company	Impact Score (Equal) $*10^6$	Impact Score (Norm.)	# of Drugs
NeuroHitech	0.59	2.33	1
GlaxoSmithKline	109.9	1.0	303
Endo International	4.1	0.93	65
AVAX Technologies	0.006	0.91	1
Acorda Therapeutics	1.7	0.88	4

Table 7.4: Companies with High Market Capitalization Normalized Impact Scores

In addition to the impact scores for the companies, we can compute the impact scores of the individual drugs. In Table 7.5, we include the impact scores of the high-impact drugs, along with the indications/diseases they are used to treat. Some of the drugs that have multiple indications approved across many countries have high impact scores. These drugs are used for neonatal disorders, musculoskeletal disorders, and malaria.

Given that drugs are used for the treatment of multiple diseases, we aggregate the impact scores over disease categories and identify those which contribute most significantly to the impact scores. In Figure 7-3, we compute the percentage of impact contributed by each disease category and find that cardiovascular diseases, cancer, musculoskeletal disorders, and diabetes are among the diseases which contribute most to the impact scores.

One potential application of the impact score is its use for impact investing (ESG investing). Investors may use these scores to find the companies with the greatest

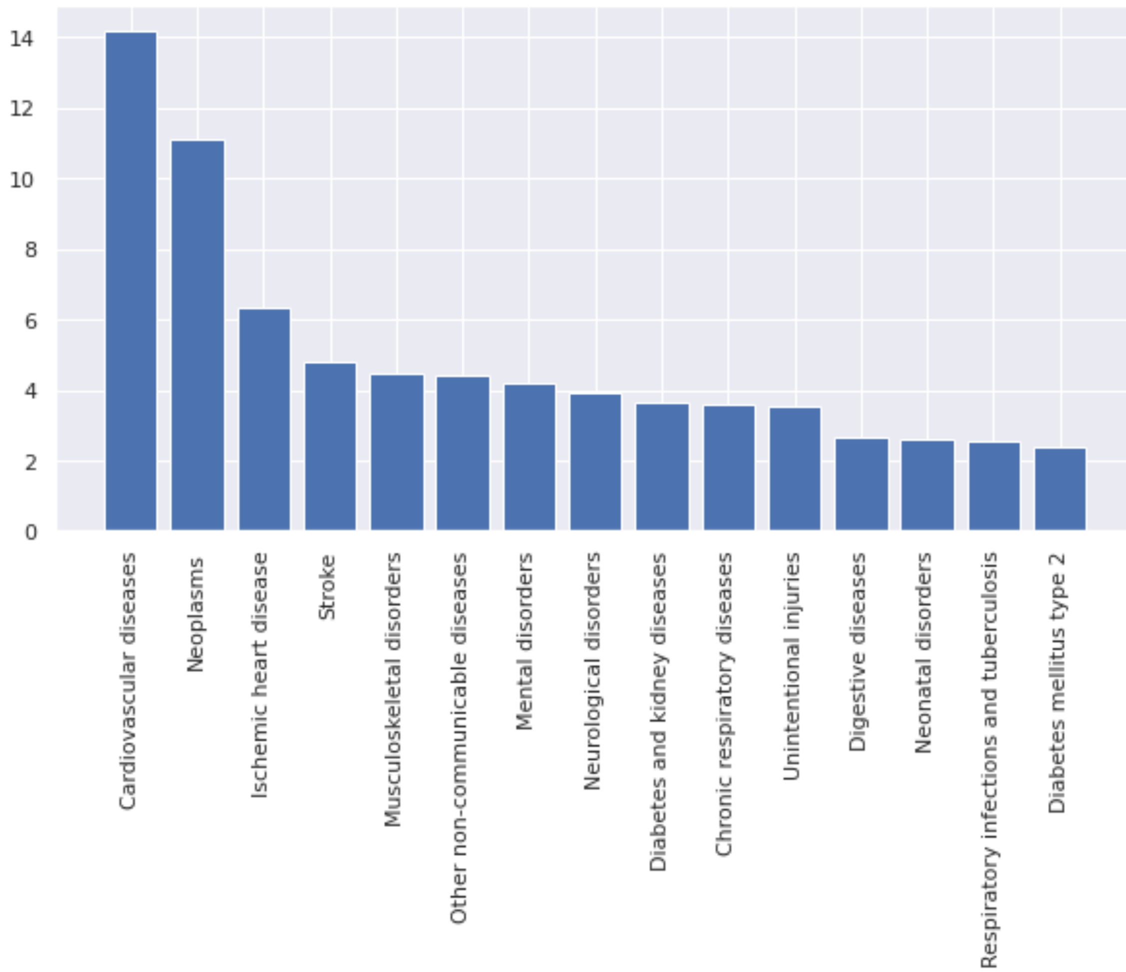


Figure 7-3: y-axis: Percentage of Impact Contributed by the Disease Category. x-axis: Disease.

Drug	Impact Score (Equal) $*10^6$	Diseases	# Country-Indications
Rho(D) Immune Globulin, CSL Limited	17.9	Neonatal disorders	55
diclofenac, Nuvo-2	13.4	Musculoskeletal disorders	3
human Rho(D) immunoglobulin, Kamada	10.5	Other non-communicable diseases, Neonatal disorders	9
diclofenac sodium gel, Novartis	10.1	Musculoskeletal disorders	10
artesunate, rectal, MMV	8.5	Neglected tropical diseases and malaria	17

Table 7.5: Drugs with High Impact Scores

impact on society. Hence, we compare these scores with the existing ESG scores released by Refinitiv. Using Refinitiv Workspace, we obtained E, S, G and ESG scores for 145 biopharmaceutical companies and compared them with our impact scores. Table 7.6 presents the correlation between the different score measures. We find that impact scores (both raw and log-transformed) are not correlated with the E, S, G, and ESG scores. However, they are negatively correlated with the market cap normalized scores due to the positive correlation of ESG scores with market capitalization. Hence, we find that existing ESG scores from Refinitiv are not correlated with our impact scores, and our scores are capturing information which is not readily available to investors.

Finally, we compute the correlation between the impact scores and historical returns of the companies. From 7.7, we find that correlation values[5] for impact scores (both raw and log-transformed) are not significant for returns computed over a small time horizon; however, they are positive and significant for a longer time horizon (2016-22). The correlation between impact scores and market cap normalized scores are sometimes negative, possibly due to the positive correlation of the returns with

[5]There is a forward-looking bias in the correlation values as impact scores are computed based on the information available on Dec-2022.

Measures	Impact Score	Impact Score (log)	Impact Score (Mkt Cap Norm)	Mkt Cap (Log)
ESG	0.31	0.04	-0.58*	0.87**
E	0.41	0.09	-0.55*	0.90**
S	0.32	0.04	-0.57*	0.86**
G	0.09	-0.07	-0.53*	0.67**
Mkt Cap (Log)	0.62**	0.17	-0.57*	1**

Table 7.6: Correlation between Scores. *: p-val < 0.05 and **: p-val < 0.01

market capitalization in our sample. If the positive correlation between long-term return and impact scores continues in the future, the portfolios based on the impact scores will achieve positive returns. Hence, along with providing high social good/impact exposure, investors may realize positive returns on their portfolios.

Return Years	Impact Score	Impact Score (log)	Impact Score (Mkt Cap Norm)	Mkt Cap (Log)
2022	-0.02	-0.02	0.03	-0.08
2021-22	0.21	0.17	-0.07	0.38
2020-22	0.03	-0.03	-0.06	0.08
2019-22	0.05	-0.11	-0.23**	0.30**
2018-22	0.13	0.04	-0.14*	0.34**
2017-22	0.13	0.12	-0.07	0.29**
2016-22	0.14*	0.15*	0.003	0.19**

Table 7.7: Correlation between Scores and Returns. *: p-val < 0.05 and **: p-val < 0.01

7.4 Conclusion

In this work, we have presented a methodology and generated scores for bio-pharmaceutical companies, which capture the impact on patients' lives due to approved drugs/treatments. We have presented the statistical properties of impact scores grouped by companies, drugs, and diseases, and demonstrated that they are not correlated with Refinitiv's

ESG scores, yet show a positive correlation with long-term historic returns. These scores can be used by investors to identify companies with significant social impact for investment purposes.

However, this work has some limitations, including several assumptions and parameters in the estimation of market share and efficacy. With improved estimates of efficacy and finer-grained data on the burden of diseases treated by the drugs, the accuracy of the impact scores could be enhanced.

Part III

Behavioural Finance

Chapter 8

Behavioural Finance: Real-time extended psychophysiological analysis of financial risk processing

In this Chapter, we study the relationships between the real-time psychophysiological activity of professional traders, their financial transactions, and market fluctuations. We collected multiple physiological signals such as heart rate, blood volume pulse, and electrodermal activity of 55 traders at a leading global financial institution during their normal working hours over a five-day period. Using their physiological measurements, we implemented a novel metric of trader's "psychophysiological activation" to capture affect such as excitement, stress and irritation. We find statistically significant relations between traders' psychophysiological activation levels and such as their financial transactions, market fluctuations, the type of financial products they traded, and their trading experience. We conducted post-measurement interviews with traders who participated in this study to obtain additional insights in the key factors driving their psychophysiological activation during financial risk processing. Our work illustrates that psychophysiological activation plays a prominent role in financial risk processing for professional traders.

8.1 Introduction

Several experimental studies in the field of behavioral economics have documented that emotional states (for example, fear, excitement, and stress) may influence the decision making of an agent [46, 116, 199]. In financial markets, sudden fluctuations in prices, changes in investor holdings, or other factors unknown to the agent may impact decisions. To better understand the biological and psychological mechanism of financial decision making under uncertainty, the emerging field of neurofinance has incorporated insights from psychology and neuroscience into theories of finance [211].

In previous studies, researchers characterized the psychophysiological (PP) activities of traders using functional magnetic resonance imaging (fMRI), hormonal level measurement, and physiological signals like heart rate, blood pressure, and skin temperature (reviewed in Section **Literature Review**). However, these techniques share the common limitation that, due to the equipment used to measure the physiological signals, the experiments are either conducted in laboratory settings rather than in the field, or the data collection procedure interferes with the day-to-day activities of traders. While controlled experimentation yields cleaner inferences, their findings lack the direct applicability to real financial settings which *in situ* experiments provide.

In this work, we conducted an *in situ* experiment and collected the physiological signals of 55 professional financial traders at a global financial institution during the entire trading day over a five-day period for each trader. Using an Empatica E4 wristband, we capture a variety of real-time physiological signals, such as heart rate, blood volume pulse, inter-beat interval, electrodermal activity, skin temperature, and three-dimensional accelerations of wrist motion, without interfering with the traders' daily routines on their trading desks. This non-invasive experimental design offers the most realistic setting to date for analyzing the traders' real-time PP activities during financial decision making.

The primary focus of our analysis was to measure the relation (if any) between a trader's PP state and their financial decisions or market events, both contemporaneously and temporally. Since a trader's PP state is not directly observed and must be inferred from the physiological signals, we measure each trader's "psychophysiological (PP) activation", defined as the collective deviation of physiological signals from their baseline values. Higher levels of PP activation indicates emotions such as excitement, stress, agitation, arousal, and irritation which elevate the physiological activities from the normal state. By analyzing the changes in the trader's PP activation levels, we identify the circumstances in which such changes coincided with, or were caused by, financial transactions or market events. The key challenge was to accurately measure the levels of PP activation, and a major contribution of our study is the novel application of the Mahalanobis distance [200] on features extracted from physiological signals as a quantitative metric of PP activation.

There are two underlying assumptions of our analysis. First, financial decision making and risk processing under uncertainty trigger PP activation in professional traders. Second, the levels of the trader's PP activation can be explained by multiple factors such as market fluctuations, the types of financial products traded, as well as the trading experience of each trader. We test four hypotheses related to these two assumptions:

- Traders monitor multiple market indices during working hours. Fluctuations in these market indices often represent trading signals to the traders and may affect the existing positions and real-time profit and loss (PnL) of the traders. We test the hypothesis that market fluctuations have causal influence on the PP activation of the traders.

- Traders who participated in our study have different lengths of trading experience ranging from 10 months to 25 years. As traders gain more trading experience, they may develop skills to manage their PP activation level when making risky financial decisions. Since we do not observe these skills, we hypothesize that the ability to manage one's activation level is positively correlated with the length of one's trading experience and test the statistical relationship between the trading experience and activation levels of the traders.

- Traders who participated in our study worked in different business divisions and traded different financial products. Since different financial products have different market volatility, dollar volume of transactions, and trading frequency, we expect the traders to have different patterns of PP activation depending on the type of their business division. We test the hypothesis that the traders' PP activation is related to their business divisions or the financial products they trade.

- As traders engage in risky financial transactions, their PP activation levels may change due to the financial decision making process before, during, and after the transaction. The PP activation levels may also depend on other factors such as the number of transactions and the average dollar volume. We examine the change in traders' PP activation before and after making a transaction and identify the period when the traders are the most highly activated. We test the hypothesis that the traders' PP activation levels are related to their trading frequency and the average dollar volume of their transactions.

We found statistically significant relationships between the levels of PP activation and financial market movements, and that different traders monitored different market signals which, in turn, had causal influence on their PP activation levels. We found that traders with more trading experience had lower levels of activation. We also found a significant relationship between a trader's activation level and the type of financial products being traded. For example, traders specializing in G10 rates, equities, commodity and foreign exchange had higher level of PP activation, while traders in securitized market generally had lower activation. Differences in the properties of the financial products lead to differences in traders' decision making process, and hence to their PP activation levels. We also found that traders on average have the highest activation levels 15 to 25 minutes after making the transaction.

We conducted follow-up interviews with 14 of the 55 traders to review their individual results, and received useful feedback which helps us interpret their PP activation patterns in the context of their specific working environments. The interviews yielded additional factors which influence traders' PP activation, such as the demand of the workday, social events, and their managerial responsibilities. We found that busier workdays and social events led to higher PP activation in some traders.

8.2 Literature review

A major challenge in conducting neurofinance field studies is the faithful measurement of the trader's real-time neural activities during live trading over extended periods of time [66, 211]. Most previous studies in the literature encountered an important tradeoff between measurement and environment, since measuring traders' neural activities in real time required a controlled experimental setting. As a result, these studies typically used mock trading and simulated market events to replace the conditions of live trading. Many studies [161, 83, 261, 115] used fMRI to continuously monitor participants' neural activity during simulated trading exercises. Another study [151] designed simulated lottery games to study the effect of elevated cortisol levels on the degree of risk aversion of the participants.

While the controlled experimental setting provides a rigorous framework for data collection and analysis, several issues prevent one from applying the methodology and conclusions of these studies to professional traders engaging in real-world trading activities. First, it is difficult to perform neural activity measurements during live trading seamlessly without interfering in the normal work routine of the trader. The act of making the measurements may have a lasting impact on the participant's mental state, which is difficult to calibrate and remove from the analysis. In addition, the neural activity profile of a participant (who may not be a professional trader) during mock trading [83, 151] is likely to be different from professional traders during live trading, since in the latter situation the traders incur real profit and loss (PnL) to their companies as a consequence of their decisions. Performing physiological measurements during live trading sessions often required non-disruptive measurement schemes at discrete points in time during or after trading hours rather than real-time monitoring. Lo *et al.* [183] used daily emotional-state surveys, and found that traders who react more intensely to their monetary gains and losses exhibited significantly worse PnL. Numerous studies [13, 67, 73, 291, 222] used saliva samples to measure traders' intra-daily cortisol and testosterone levels and studied the relation between traders' hormone levels and their trading performance measured by PnL. Although saliva sampling may faithfully reflect a trader's PP state during live trading without causing much disruption to their normal work routine, it is difficult to perform such measurements continuously in time, and the studies cited above collected saliva samples twice or three times each day. These measurement schemes cannot effectively capture a trader's real-time PP responses to transient market events and fluctuations in asset prices and market indices.

Lo and Repin [182] were the first to measure traders' real-time physiology during live trading sessions. They also collected financial market data relevant to the traders' financial decisions synchronously during the physiology measurements. The time series data allowed them to identify PP responses to transient market events and periods of high market volatility, though their study involved relatively few subjects (10 traders) and short periods of measurement (ranging from 49 to 83 minutes) when the physiological signals were collected from each trader.

In this study, we extended the line of research of [182] and collected physiological signals of 55 professional traders at a global financial institution. Data collection for each trader lasted five consecutive workdays over a week, and six to eight hours each day. The larger sample size and longer measurement periods allowed us to study the PP behavior of traders in a more unconstrained and natural setting. We used the Empatica E4 wristband to collect high-frequency physiology time series data without disrupting the traders' normal work routine. We also collected relevant market data and financial transaction data of individual traders to analyze their PP responses to market fluctuations and financial transactions.

Another contribution of this study is our novel method for measuring the PP activation from physiological signals to capture psychological states such as excitement, stress and irritation. Higher activation is known to trigger psychological, behavioral, and physiological changes in the brain [207]. These changes influence different biological signals, such as electrocardiogram readings, blood volume pulse, electrodermal activity, heart rate variability, and skin temperature [120]. Since stress is an important cause of PP activation, the rich literature of stress detection from physiological signals may be applied to measure PP activation. Al-Shargie *et al.* [7] measured electroencephalography and functional near-infrared spectroscopy on the prefrontal cortex for stress assessment. Several studies [252, 167, 163, 49] used features extracted from electrodermal activity to train stress detection machine learning models, while others [167, 70, 241, 156] analyzed heart rate variability for stress detection. In addition, photoplethysmography [62] and skin temperature [281] are also used for stress detection. Fernandez *et al.* [108] proposed a system which uses commercial devices to detect stress and alert in traders in real time, which is of direct interest to our purpose.

Previous experimental studies generally used artificial methods to elevate the activation levels in the subjects, and measured their biochemical markers or physiological responses synchronously [20]. This "ground truth" information of the activation patterns makes it possible to train machine learning models for PP activation analysis. However, such "ground truth" information is not available in our study since we are agnostic of the factors which elevate the traders' activation. Instead, we propose a novel metric for PP activation using the Mahalanobis distance [200] of features extracted from physiological signals, which to the best of our knowledge has not been previously explored. This metric allows us to rigorously test the statistical relationship between PP activation and various aspects of financial decision making and risk processing, discussed in the next section.

8.3 Materials and methods

To rigorously analyze how financial risk processing affects a trader's PP state, we collected both financial transaction records and the real-time physiological signals (Table F.1 in Appendix) of 55 professional day traders at a global financial institution during their normal trading activities. Before data collection, we collected the demographic information of the traders such as gender, trading experience, and business division via surveys. To understand the impact of market fluctuations on the trader's PP activation, we also acquired 10 intraday financial market time series (Table F.2 in Appendix) commonly monitored by the traders during the same period as we measured their physiology. The details of data collection and preprocessing procedures are described in Appendix.

Measuring PP activation

From the physiological signals of the traders, we constructed a synthetic metric to characterize each trader's PP activation, defined as the physiological response to emotions like excitement, stress, and irritation. We chose to measure PP activation since it is ubiquitous in the context of financial risk processing. Events such as high market volatility, sudden gain or loss in the trader's portfolio, or receiving instructions from clients to place transactions may induce elevated activation in traders. Previous studies showed that a high level of activation is associated with reduced performance [140, 229]. In addition, prolonged periods of high activation have an adverse impact on health [205, 65]. We extract PP activation from the physiological features in the following way:

- Heart Rate Variability (HRV): HRV measures the variations in the beat-to-beat intervals. Inspired by [156], we calculated four features from the raw HRV signal to measure activation: mHR, the average number of beats per minute; mRRi, the average value of the interbeat (RR) interval; SDNN, the standard deviation of the RR interval; and RMSSD, the root mean square sum of the successive interbeat interval difference. All features are calculated as time series using a 5-minute trailing window, similar to [156].

- Electrodermal Activity (EDA): EDA measures the electrical conductivity of the skin. Elevated activation increases the skin temperature and sweating, which in turn changes the EDA. We calculated three features from the raw EDA signal to measure activation: mAmp, the average of the raw skin conductance signal; Slope, the average of the absolute first difference of the skin conductance signal;

and Events, the number of times the EDA signal increases more than $0.05\mu S$ in less than 5 seconds. These features are inspired by [163].

- Blood Volume Pulse (BVP): BVP is measured by a photoplethysmogram (PPG) sensor, which uses optical technique to measure the blood flow rate and blood volume. Blood pressure is then estimated using BVP readings [293]. Elevated cardiovascular activity due to higher activation changes the BVP. We calculated three features from the raw BVP signal to measure activation: the average of the absolute value of the BVP; the minimum value of BVP, and the maximum value of BVP. These features are inspired by [163].

- Skin Temperature (TEMP): TEMP generally increases during periods of high activation. The changes in skin temperature depend on the specific body region being measured [281] and changes in room temperature. We assume that changes in room temperature occur much more slowly than the sudden onset of events that lead to high activation. We compute the instantaneous rate of change in TEMP to capture the change in levels of PP activation.

Once we extracted these features as a column vector $y_t \in \mathbb{R}^d$ at each time t for a trader, we computed the aggregate measure of PP activation using the Mahalanobis distance d_M [200],

$$d_M(y_t) = \sqrt{(y_t - \mu_t)^T \Sigma_t^{-1} (y_t - \mu_t)} \tag{8.1}$$

where the column vector $\mu_t \in \mathbb{R}^d$ and matrix $\Sigma_t \in \mathbb{R}^{d \times d}$ are the rolling sample mean and covariance matrix computed using $y_1, \ldots, y_t$. The superscript T denotes vector transposition. We use the rolling mean and covariance to prevent the look-ahead bias of computing PP activation at time t using signals measured in future time $t' > t$. The Mahalanobis distance captures the trader's PP activation as the collective deviation of physiological activities y_t from their baseline state μ_t and adjusts for the covariance between the components of y_t, which tend to be highly correlated if they are extracted from the same underlying physiological signal measured by the wearable device (shown in Table F.6 of Appendix).

Using the Mahalanobis distance, we computed five different measures of PP activation: the overall PP activation (using features extracted from all physiological signals), and the individual HRV, EDA, BVP, and TEMP activation (using features extracted from one physiological signal) for every trader. In the subsequent analysis, we focus on the overall PP activation since it captures the collective excitation of all physiological activities simultaneously.

Next, we identify the time periods when each trader is under mild and extreme levels of PP activation by measuring their deviations from mean. We label the trader as "mildly activated" at time t if her PP activation $d_M(y_t)$ satisfies $d_M(y_t) > \mu + 1.5\sigma$ where μ and σ denote the mean and standard deviation of her PP activation during the trading day. Details of computing μ and σ are provided in Section F.6 in Appendix. Similarly, we label the trader as "extremely activated" at time t if $d_M(y_t) > \mu + 3\sigma$. We choose thresholds 1.5 and 3 based on qualitative assessment of the magnitude of mild and extreme activation. Future extensions of our work may use data-driven methods to identify periods of mild and extreme activation for each trader.

Finally, we use three aggregate metrics to characterize the PP activation patterns of each trader: activation proportion, activation length, and average activation. Activation proportion measures the percentage of time when the trader was at a mild or extreme level of activation during a trading day. Traders with higher activation proportion tended to be more susceptible to elevated activation. Activation length measures the average duration of the mild or extreme activation periods of a trader. Traders with longer activation length tended to remain activated for longer periods of time after the sudden onset of events which triggered activation. Average activation measures the average value of PP activation over the trading day and reflects both activation proportion and length. Traders with higher activation proportion or length also have higher average activation. These aggregate measures allowed us to directly compare the PP activation patterns across different traders and trading days. The definition and computation of activation proportion and activation length are described in Section F.6 in Appendix.

Activation level attribution regression

To identify the relationship between individual factors and the PP activation of the traders, we perform the following regression:

$$y_{i,t} = \alpha + \beta * bis_i + \gamma * gen_i + \delta * amt_{i,t} + \theta * exp_i + \zeta * numT_{i,t} + \eta * vol_{i,t} \quad (8.2)$$

where $y_{i,t}$ denotes the activation metric of the trader i on day t, bis_i the business division (one-hot encoded), gen_i the gender (male/female), $amt_{i,t}$ the average dollar value of the transactions by trader i on day t, and $numT_{i,t}$ the number of transactions by the trader on day t. Since traders' activation may also be correlated with market volatility on different days, we also include $vol_{i,t}$, the vector of volatility of different market indices on day t. The right-hand-side explanatory variables of the regression were extracted from trader survey, financial market indices, and traders' transaction data. We use the average activation as the left-hand-side dependent variable $y_{i,t}$ since it reflects the other activation metrics (proportion and length).

Granger causality test

We test whether there is Granger causality relation between fluctuations of market indices and trader's PP activation. Specifically, let y_t and m_t denote the minute-level time series of a trader's PP activation and a market index on a trading day. Since the activation time series is measured with 1Hz (higher frequency than market indices), we perform frequency averaging on y_t to get the minute-level activation time series $\overline{y}_t$. To account for non-stationarity in the time series, we perform the Granger causality test on the first difference $\Delta\overline{y}_t = \overline{y}_t - \overline{y}_{t-1}$ and $\Delta m_t = m_t - m_{t-1}$. For each trader and each market index on a given trading day, we test the null hypothesis that Δm_t does not Granger-cause $\Delta\overline{y}_t$. The Granger causality test is performed with the sum of squared residual (SSR) test with chi-squared distribution using the **statsmodels** package in the SciPy library and a lag of 10 minutes. To correct for multiple hypothesis testing across all traders, we perform the Holm-Bonferroni correction for each market index with significance level 0.05.

Activation around financial transactions

We analyzed the pattern of trader's PP activation around financial transactions. Specifically, we computed the evolution of PP activation using a rolling window before and after placing a trade in the following way:

- We first calculated the z-scores of PP activation z_t so that the activation levels may be consistently compared across different time periods and traders.

- Next, for a transaction at time l placed by the trader, we calculated the mean values of z_t in the windows $[z_{l-30min}, z_{l-25min}]$, $[z_{l-25min}, z_{l-20min}]$, ..., $[z_{l+20min}, z_{l+25min}]$, $[z_{l+25min}, z_{l+30min}]$. This computed the time series of activation averaged over 5 minute windows in 30 minutes timeframe before and after performing the transaction.

- Finally, for each trader in a trading day, we averaged the rolling window calculation results across all transactions made by this trader on this day. We computed the evolution of PP activation around a transaction by averaging the results across all traders and trading days.

Individual trader meetings

The 55 traders who participated in our study had diverse levels of trading experience and traded a variety of financial products. Since the physiological signals were recorded during their normal working hours, their PP activation levels may have been affected by idiosyncratic sources that were exogenous to their trading activities and unknown to us. To investigate these idiosyncratic sources of PP activation and relate them to our findings, we invited all participating traders to meet individually with us after we completed the PP activation analysis. Due to traders' availability, 14 of 55 traders participated in the individual meetings. During each meeting, we presented the individual findings to the trader and received the feedback from the trader. We applied the insights from these trader meetings to identify additional sources of PP activation.

Metric	Min.	Max.	Mean	Std. Dev.
Average Activation	2.76	5.34	3.79	0.37
Mild Activation Proportion (%)	7.42	17.57	11.54	1.99
Extreme Activation Proportion (%)	0.45	6.78	2.71	1.12
Activation Length (s)	26	1031	152	94

Table 8.1: Summary statistics of four PP activation metrics.
Average activation is a dimensionless quantity. Mild and extreme activation proportions are measured in percentages (%). Activation length is measured in seconds (s).

Ethics Statement

The study was approved by the following IRB/ethics committee: MIT Committee on the Use of Humans as Experimental Subjects (COUHES). The study approval number is 0403000144 and the written consent was obtained prior to the study.

8.4 Results

We present the analysis results of traders' PP activation and investigate the different sources that may cause elevations in PP activation levels. We discuss both general patterns of trader's PP activation observed across all traders and idiosyncratic characteristics specific to individual traders.

Variation in trader's activation levels

We analyzed the variation in PP activation patterns across different traders as reflected by four activation metrics: average activation, mild activation proportion, extreme activation proportion, and activation length. Figure 8-1 shows the histograms of these metrics for all traders. A trader's PP activation metric on each trading day is counted separately in the histogram to account for inter-day variation and there are a total of 242 trader-day pairs in each histogram. The summary statistics of each histogram is shown in Table 8.1.

We observed significant variation in all PP activation metrics across the traders. Several traders were at mild activation for 17% of working hours on certain days,

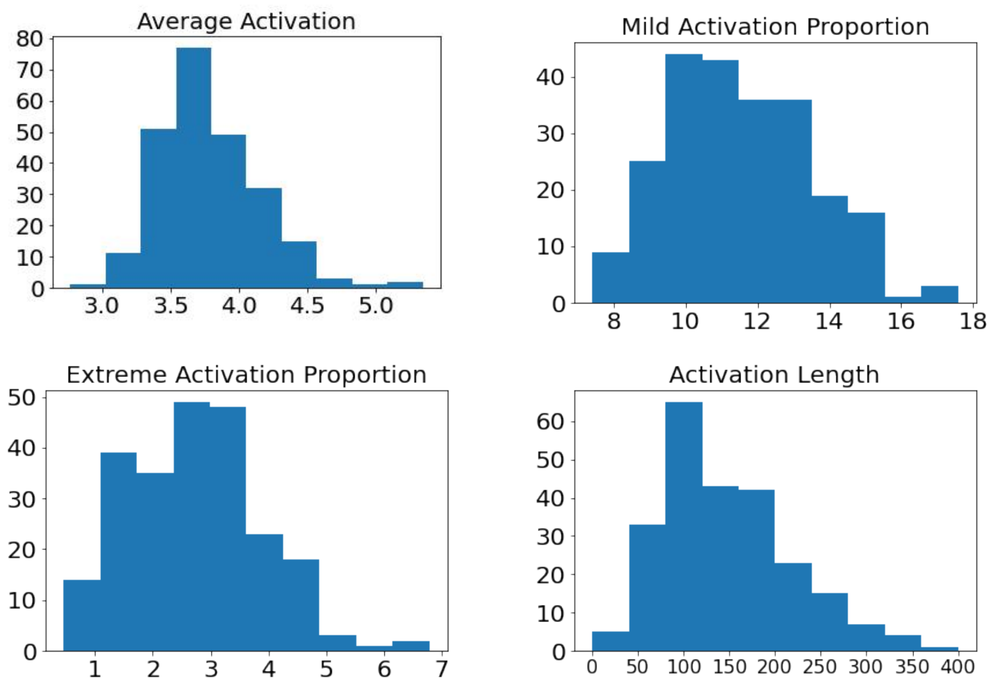


Figure 8-1: Histograms of four PP activation metrics.
The y-axis represents the number of trader-days and the x-axis represents the activation metric values. We observe considerable variations in PP activation levels across traders. Average activation is a dimensionless quantity. Mild and extreme activation proportions are measured in percentages (%). Activation length is measured in seconds (s).

while others were at mild activation for less than 8% of the time. Similarly, some traders returned to their normal activation levels in less than 30 seconds from the onset of elevated activation, while others took as long as 17 minutes. The significant variation in PP activation metrics can partly be ascribed to the diverse characteristics of the traders in various aspects, which motivates us to investigate the systematic and idiosyncratic factors which lead to the variation in PP activation levels.

Factors influencing PP activation

We performed the regression (Eq:8.2) to identify the relationship between different factors and the PP activation of traders. The dependent variable of the regression $y_{i,t}$ is the average PP activation of trader i on day t. The correlations between the covariates of the regression are summarized in Table F.5 in Appendix.

Gender

We do not find statistically significant relationship between the average PP activation and the gender of a trader (difference between the regression coefficients of male vs. female: -0.02, p-value: 0.83).

Trading experience

We find that traders with more trading experience have lower average PP activation (coefficient: -0.02, p-value: 0.01). The possible explanation is two-fold: traders may become more adept at making risky financial decisions under uncertainty with little emotional fluctuations or may acquire certain skills over time to actively manage their PP activation levels during financial risk processing.

Dollar volume, number of transactions, and market volatility

The average dollar volume and number of transactions have positive regression coefficients 0.02 (p-value: 0.26) and $0.4e-4$ (p-value: 0.47), respectively. While we observed an overall positive correlation between the dollar volume, number of transactions and the average PP activation of the traders, additional factors of their trading activities might further influence whether the transaction causes high or low PP activation and lead to the large p-values. During individual trader interviews, we found that traders' PP activation highly depends on type of transaction (via a client's order or the trader's own decision) and the changes in risk exposure as a result of the transaction. Similarly, the regression coefficients of volatility of different market indices are not statistically significant. This motivates us to analyze the time series characteristics of market fluctuations and traders' PP activation via the Granger causality test discussed later.

Other factors

We also found statistically significant relations between the trader's average PP activation and the type of financial products traded. A detailed discussion of these results is provided in Section F.8 in Appendix. During individual trader meetings, traders reported additional idiosyncratic factors that influenced their PP activation, including whether they had a busy schedule, their managerial responsibilities, and even the anticipation of the social events after work. These idiosyncratic factors were not measured in our experiment and may contribute to the variation in traders' PP activation unexplained by the regression.

Market fluctuations have causal impacts on traders' PP activation

Typically, each trader manages a portfolio of many active positions and monitors a variety of market indices and events during the trading hours. It is reasonable to expect that increased volatility in market indices relevant to their portfolio values may cause the traders' PP activation levels to become elevated. We tested the null hypothesis that market fluctuations do not have causal influence on traders' PP activation via the Granger causality test, with Holm-Bonferroni correction at 5% significance level.

To test whether the observed Granger causality relations between market index

Market Index	**Number of significant GC relations**
Credit-Default Swap Index IG	22
Credit-Default Swap Index HY	19
USD Index	3
10Y US Treasury Future Price	3
S&P 500 E-mini Future Price	4
5Y US Treasury Future Price	6
Crude Oil Futures Price	4
VIX Futures Price	7

Table 8.2: Granger Causality Test
The number of statistically significant Granger causality relations between each market index and traders' PP activation, with Holm-Bonferroni correction at 5% significance level.

and PP activation are statistically significant, we plot the distribution of p-values for each pair of PP activation and market index in Figure F-4 in Appendix. Under the null hypothesis H_0 that the market index does not Granger-cause traders' PP activation, the distribution of p-values of Granger tests should follow a uniform distribution on $[0, 1]$. We perform a one-sample Kolmogorov–Smirnov test and reject the null hypothesis H_0 for six of the eight market indices (Credit Default Swap Index Investment Grade (IG), Credit Default Swap Index High Yield (HY), S&P 500 E-mini Futures Price, 10Y US Treasury Futures Price, 5Y US Treasury Futures Price and Crude Oil Futures Price) at significance level 0.05.

The number of statistically significant Granger causality relations between market fluctuations and traders' PP activation is shown in Table 8.2. We observe that the two credit default swap (CDS) indices were the most common sources of elevated PP activation for the traders. Since CDS contracts act as insurance against credit defaults, it is reasonable that fluctuations in CDS indices have systematic impact on multiple financial products and elevate the PP activation levels among the largest number of traders.

Financial transactions elevate traders' PP activation

Traders regularly engage in financial transactions which induce considerable risks to their PnL performance. Real-time financial risk processing during a transaction can be a major source of high PP activation for the traders. Since transactions are placed at irregular points in time, we performed an event study to investigate the changes in trader's PP activation levels before and after a transaction, as described in **Methods and Materials** Section.

Figure 8-2 shows the evolution of the average PP activation levels around a transaction, where the results are averaged across all traders. The vertical axis represents the average PP activation and horizontal axis indicates five-minute intervals around the transaction, where the transaction time is defined at $t = 0$. The left half of the figure corresponds to PP activation levels before the transaction, and the right half to those after. To account for the variation across traders, we also plot the 95% confidence band of PP activation.

We observe that traders tended to have the highest PP activation in the time window of 15 to 25 minutes after the transaction. We also observe a local peak of high PP activation 5 minutes prior to the transaction. During individual trader interviews, traders pointed out that their PP activation levels depended on the risk exposure of the transaction, as well as whether the transaction was placed via a client's order or per the trader's own discretion. In addition, traders mentioned that it usually takes 15 to 30 minutes after a transaction to confirm whether the transaction is executed as intended and to observe its effect on the trader's PnL. This anticipation effect partly explains the elevated PP activation levels 15 to 25 minutes after transaction, shown in Figure 8-2.

8.5 Discussion

Our analysis confirms that affect, as evidenced by the fluctuations of PP activation due to market fluctuations or financial transactions, plays a prominent role in the financial risk processing and decision making process for professional traders in their

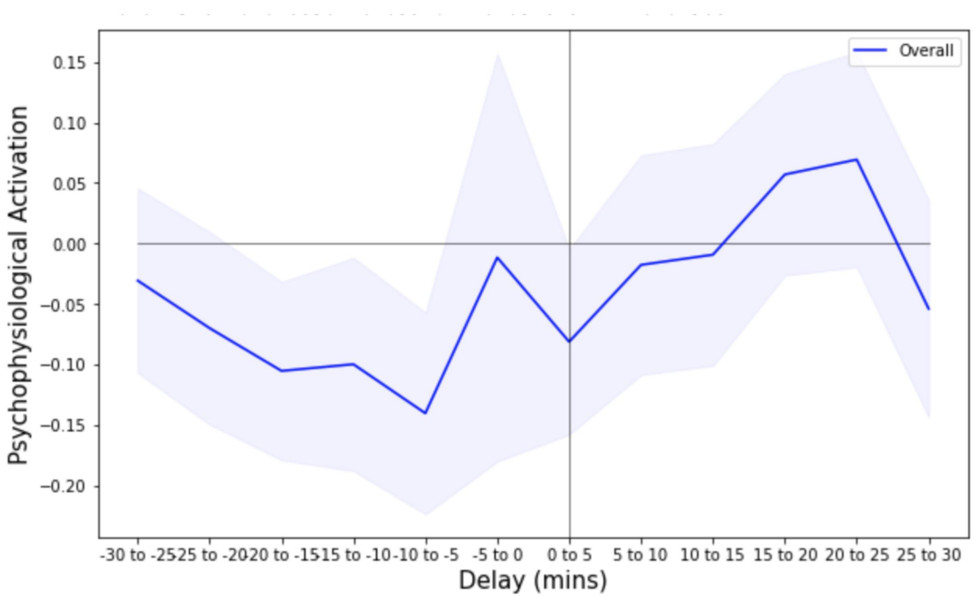


Figure 8-2: Evolution of trader's PP activation around a transaction.
The shaded region corresponds to the 95% confidence band. The x-axis is the time difference (in minutes) between the transaction time, defined as time 0, and measurement time. The y-axis is the average PP activation value. We observe that traders have the highest PP activation during 15 to 25 minutes after the transaction, as well as a local peak 5 minutes before the transaction.

daily trading activities [182]. Our results provide contextual support for the growing literature of affective science investigating the relations between emotions and decision making [187, 170, 71] in the case of professional traders, whose profession requires them to make rational decisions under large uncertainties and maximize their payoffs.

Our study also illustrates the feasibility of conducting large field experiments in affective sciences with non-disruptive physiological measurement and the rich insights one can extract from this data using appropriate statistical techniques. For example, by utilizing the time series properties of PP activation, we overcome the challenge of performing causal inferences with observational data and demonstrate the causal relationship between certain market indices and traders' PP activation. The novel use of Mahalanobis distance to capture the aggregate PP activation from individual physiological signals may be widely applied in many other contexts in this field.

Our analysis has several limitations to be addressed in future works. First, while we observed the presence of affect in financial risk processing, our analysis does not distinguish between positive and negative affects or investigate whether affect constitutes a beneficial or adverse driver of the traders' financial performance. Previous studies revealed that positive affect leads to improved decision making [147] while negative affect such as fear and anger distorts the perception of risk [169] and causes myopic decisions [171]. During individual trader meetings, 12 of the 14 traders pointed out that they typically experienced high levels of activation when their real-time trading performance (measured by PnL) exhibited losses or fluctuations. Future study may analyze the relation between traders' PnL performance and their PP activation under various market conditions using appropriate statistical or machine learning techniques [56]. Using such findings, one may even quantify the impact of affect on the trader's PnL and prescribe strategies for traders to actively manage their affect and improve their trading performance.

In addition, while the field experiment design allows us to faithfully measure the physiological and affective state of professional traders in their natural working environment, it also prevents us from observing and controlling all the factors which influence the trader's affect. As a result, our analysis does not make the important distinction between integral emotion (the emotion induced by financial risk processing which is our main interest) and incidental emotion (the emotion carried over from non-trading situations such as performing managerial tasks or participating in work meetings) [170]. This limits the interpretation of our findings since previous studies showed that incidental emotions affect risk perception [249] and induce bias on the financial decisions [97]. Future studies may improve the experimental design by asking participants to wear the measurement device only during trading activities or report the time periods when they are mainly occupied with non-trading activities. While this improves the statistical inference, researchers must also ensure that participants adhere to the experiment protocols.

Finally, from an evolutionary biology perspective, professional traders must adapt to their highly competitive, dynamic and uncertain working environment where significant portions of their compensation depend on their PnL. It is natural to hypothesize that the biological and psychological mechanisms of financial risk processing are markedly different between professional traders and the rest of the population [179]. Another interesting extension of our work is to conduct a similar experiment with a control group of subjects with little or no trading experience. In this way, one can determine the unique characteristics of professional traders' cognitive and emotional faculties which enable them to thrive in the highly dynamic and uncertain working environment.

8.6 Conclusion

We conducted a large field experiment and measured the real-time physiological signals of 55 professional traders during their normal trading activities over a five-day period. Using a novel metric of PP activation based on the Mahalanobis distance, we found large variations in PP activation across traders. We showed that the PP activation is correlated with and influenced by multiple factors, such as market fluctuations, financial transactions, as well as trader's experience and types of financial products traded. Our analysis confirms the prominent role of affect in financial risk processing for professional traders. Future studies may analyze the impact of traders' affect on their trading performance and the difference in affect between professional traders and subjects with no trading experience when making risky decisions under uncertainty.

Part IV

Explainable Artificial Intelligence (xAI)

Chapter 9

xAI : Explainable Machine Learning Models of Consumer Credit Risk

In this Chapter, we create machine learning (ML) models to forecast home equity credit risk for individuals using a real-world dataset, and demonstrate methods to explain the output of these ML models to make them more accessible to the end user. We analyze the explainability for various stakeholders: loan companies, regulators, loan applicants, and data scientists, incorporating their different requirements with respect to explanations. For loan companies, we generate explanations for every model prediction of creditworthiness. For regulators, we perform a stress test for extreme scenarios. For loan applicants, we generate diverse counterfactuals to guide them with steps towards a favorable classification from the model. Finally, for data scientists, we generate simple rules that accurately explain 70-72% of the dataset. Our study provides a synthesized ML explanation framework for all stakeholders, and is intended to accelerate the adoption of ML techniques in domains that would benefit from explanations of their predictions.

9.1 Introduction

The total U.S. household debt at the end of the fourth quarter of 2020 is estimated to be $14.56 trillion, with 189.6 million new credit accounts opened within the prior 12 months.[1] The total amount of home equity lines of credit (HELOCs) at the same time is around $300 billion.[2] Although small in relative terms compared to the total U.S. household debt, this number is still significant economically, and HELOCs currently account for the majority of the banking industry's home equity portfolios. Given this massive amount of household debt, even a low delinquency rate can significantly affect the operation of the business and the financial system as a whole. This potential impact makes the study of credit default risk an important real-world task.

Lenders generally use consumer credit ratings to grant and structure the terms of credit to consumers. To compute these credit ratings, a variety of factors that gauge the creditworthiness of individuals have been described in the literature, which extends as far back as the 1940s [54]. One popular example is the FICO scores created by the Fair Isaac Corporation, which are used by over 90% of top lenders when making lending decisions.

Recent advances in computing, innovative algorithms, and an explosion in the quantity of data have contributed to the growing success of complex nonlinear machine learning models, including what are known as "deep" learning models. Unlike traditional machine learning techniques such as logistic regression and decision trees, deep learning models may have an extraordinary number of parameters. They are able to automatically learn nonlinear representations and interactions of input features from large datasets, a process that helps them to achieve superior performance compared to other machine learning methods. ML models have been explored for diverse applications in economics and finance [127]. In particular, they are widely used for a variety of different credit risk applications, including peer-to-peer lending [197, 93], mortgage risk [246, 164, 60], credit card risk [45], consumer credit risk [148], fair credit allocation [269], sovereign default risk [87], and a large pool of loans [260].

However, regulatory compliance is a major obstacle in the adoption of black-box models for credit risk modeling. In the United States, the Fair Credit Reporting Act of 1970 mandates that lenders must be able to disclose up to four key factors that

[1] Source: New York Fed Consumer Credit Panel/Equifax

[2] Source: FRED Economic Data.

adversely affected the credit score of a rejected consumer. More recently, the European Union's General Data Protect Regulation of 2018 creates a right to explanation, whereby users may ask for an explanation of an algorithmic decision that was made about them [130]. Similarly, the European Artificial Intelligence act of 2021 imposes the requirement of global explainability for artificial intelligence algorithms.

Apart from regulatory compliance, different stakeholders have different requirements for explanation and transparency. For example, a loan applicant may want to know the possible steps that she could follow to make her creditworthy. Similarly, loan companies may need to provide explanations for their decisions regarding the creditworthiness of applicants, while system developers may need to understand the specific features and relationships that underpin their models.

In this chapter, we first model a credit risk forecast on a real-world home equity line of credit (HELOC) dataset released by the Fair Isaac Corporation (FICO). This dataset is widely used to study interpretability (see, for example, [244] and [59]). We developed a wide range of ML models, including interpretable rule-based models (e.g., inductive logic programming and optimal trees) and black-box ML models (e.g., neural networks and random forests). We compare these models and find that neural networks outperform other linear and nonlinear models, reaching an accuracy of 74.75% on a task to predict the creditworthiness label provided by FICO. We also find that properly-constructed explainable rules can reach an accuracy of up to 70-72%. Our focus in this paper is the *explainability* of ML models. As such, we do not view these models as optimal in terms of prediction accuracy, and they are constructed as proofs-of-concept for studying the explainability of black-box models.

The *explainability* of a model means the ability to give answers to the different stakeholders involved in the decision-making process [72]. Interpretability for one party need not be translated to interpretability in the eye of another party. Hence, in this paper, we identify the different stakeholders involved in credit risk management—loan companies, regulators, loan applicants, and data scientists—and provide explanations of the models according to their needs.

For loan companies, we generate interpretable explanations for every prediction using two post hoc explainability tools, Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), that are able to provide reasons for loan denial. We modify the methodology of the LIME algorithm to incorporate the covariance structure of different features implied by constraints on their relationship. The ability to generate explanations makes a model transparent about the relationships it has learned. For example, we identify that having a higher average age of line of credit leads to a decrease in the probability of default. In addition, these explanations also make it possible to find representative borrowers from the past to aid loan officers in human decision-making.

For regulators, we suggest a method of investigating fairness of these models. In addition, regulators may also be interested in evaluating the model's functionality in extreme scenarios. In anticipation of this issue, we provide a methodology to perform stress tests to help the regulators analyze the model's behavior in extreme scenarios. In practice, the threshold for regulators to adopt any model will be high and requires a thorough investigation into the above issues using a comprehensive dataset. Our study would not be sufficient to definitively answer such questions given the limited data available publicly. Nonetheless, we provide a methodology to prove the concept and demonstrate the possibility to carry out such analysis in a systematic way.

For loan applicants, because they are interested in getting approval for a loan, we generate a diverse set of suggestions that can help individuals deemed non-creditworthy become creditworthy, even if the black-box ML models are used for loan approval decisions. For individuals who already have been approved for loans, we are able to generate suggestions that can help them remain creditworthy. These counterfactual suggestions are generated by incorporating constraints based on specific items of domain knowledge, which makes these suggestions more practical in real-world applications. We generated suggestions successfully for 99% of the individuals in our testing dataset.

Finally, for data scientists, we are able to summarize the dataset using a few simple rules that can help them understand the relationships and structure in the dataset. Understanding these relationships may provide a data scientist with the insights to develop better models in the future. Using inductive logic programming, we are able to explain 70% of the dataset using a single rule, and up to 72% using two simple rules.

The key contributions of this paper can be summarized as follows:

- We identify and summarize the explainability requirements for different stakeholders involved in credit risk management.

- We identify the right tools from the literature that can help in answering necessary explainability questions for different stakeholders in a systematic way.

- We modify the existing explainability tools and generate better explanations for different stakeholders.

This chapter demonstrates that the functionality of black-box ML models can be explained to a range of different stakeholders, if the right tools are applied to the task. These progresses unlock the future potential of applying AI to improve credit modeling. New explainable AI tools can help stakeholders make counterfactual predictions and explain a model's output, leading to the ability to answer many "what-if" questions. This includes stress tests for extreme scenarios, and informing applicants about how to improve their creditworthiness. More generally, it demonstrates the importance and potential of explainable AI to advance classical tasks in traditional fields like credit modeling. However, we emphasize that the data used in the study and the models developed are meant to be proofs-of-concept. But, similar analysis can be performed for other larger datasets if they are made available.

The remainder of this paper is structured as follows. Section 9.2 discusses the relevant literature. Section 9.3 describes the different models used in our paper, and Section 9.4 describes the dataset. Section 9.5 evaluates the prediction performance across different models. Section 9.6 develops a range of explainability methods for various stakeholders. Section 9.7 summarizes our findings and concludes the paper.

9.2 Literature Review

A range of statistical and operational research methods has been used over the long history of credit scoring models [275]. Many existing models learn relationships between risk factors and borrower behavior, either using logistic regression [48, 74, 99, 2] or cox proportional hazard model [132, 84, 265]. Some studies find non-linear relationships between the risk factors and mortgage risk [99, 113, 3].

More recently, machine learning techniques have been adopted for credit risk models. For example, [246] develop a deep learning model and find significant nonlinearities in mortgage delinquency, foreclosure, and prepayment risk. [164] predict mortgage default using convolutional neural networks. [172], [275], [39], and [127] give detailed surveys of ML methods for credit risk forecasting. In the meantime, [127] also highlight the shortcomings of ML with regard to the explainability of the models for fintech applications.

On the other hand, an extensive literature has been developed about explainable AI in healthcare, computer vision, and natural language processing [214, 1]. Even though these techniques were developed for other domains, some have been adopted for explaining credit risk forecast models.

A variety of solutions have been proposed to deal with the shortcomings of ML models used in credit risk forecasting. Interpretable ML models have been used by [155] (decision trees for a consumer credit risk model), and [223] (a multiclass rule-based model for corporate credit ratings). Similarly, [94] propose an interpretable penalized logistic tree regression model for credit scoring, while [59] use an interpretable two-layer additive risk model for a home equity line of credit dataset. Some studies have explored the use of network-based [123] or latent factor [6] models for peer-to-peer lending, which can be more interpretable than traditional deep learning models. A two-layer additive risk model is the award-winning model in the recent FICO data science challenge. For these models, interpretability generally comes at the cost of performance compared to black-box models. For example, [50] found that random forests outperform decision tree classifiers. In this paper, we show that a neural network outperforms the two-layer additive risk model proposed in [59], which we use as a benchmark. Most of these explainable ML models are developed for applications involving structured data.

There is also a rapidly growing literature on post hoc interpretable ML methods. These methods can generally be used on both structured and unstructured data. [246] use first-order derivatives and cross partial derivatives of the fitted transition probability with respect to features to study marginal and interaction effects. [44] use Shapley values to explain tree-based ensemble models and apply correlation networks to group the borrowing companies using the derived explanations. Similarly, [134] use Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to obtain local and global explanations for models trained on a Lending Club dataset. [8] propose a deep learning-based approach that combines the outputs of tree-based ensemble models and neural networks to predict consumer default, and use SHAP to provide model explanations. Other methods that use SHAP for explaining credit risk models include [15] and [38]. More recently, [235] propose a human-in-the-loop ML method that combines a post hoc explanation from SHAP with explanations from credit risk experts. [244] propose a minimum set cover problem to generate a rule-based summary of a machine learning model on the home equity line of credit dataset. Similarly, [206] propose a rule extraction method from a trained support vector machine (SVM) for credit scoring. There are some post-hoc interpretability methods, including Accumulated Local Effects (ALE) and Leave One Covariate Out (LOCO) [168, 14], that can be used for finding the global relationship between input features and output. Finally, [232] propose the layerwise relevance propagation and activation analysis of hidden units of a neural network.[3]

While these methods generate explanations of model outcomes, they are not tailored specifically for the different stakeholders involved in the credit risk pipeline. For example, many of the above-mentioned methods do not provide local explainability that is necessary for many real-world applications. Because interpretability for one

[3] There are also other interpretability methods, such as those discussed in [125] and [124], that are yet to be applied to credit risk management.

party may not always translate to interpretability in the eye of another party, none of the existing methods can be generalized for all stakeholders. In addition, they do not embed the specific constraints of the dataset in their methodology of generating explanations, which may lead to explanations that are not practical or useful. Interpretability for different stakeholders has been introduced in a demo by IBM,[4] but it remains incomplete and does not use the state-of-the-art methods that we employ in this study.

In this paper, we present a methodology that generates explanations of black-box models for different stakeholders, a direction that has not been well explored previously. In some cases, we improve existing methods by making necessary modifications for our purposes, while in others we develop new models. In the process of generating these explanations, we embed constraints to ensure that explanations and suggestions are meaningful and can be adopted for real-world use.

9.3 Machine Learning Methods

The fundamental goal of credit scoring is to determine the creditworthiness of an individual. Simply put, it is a binary classification task that labels credit applicants as creditworthy or non-creditworthy. A creditworthy applicant is likely to repay their financial obligation, while a non-creditworthy applicant is not.

We frame the consumer credit risk classification problem as predicting the probability of default for a borrower. More specifically, given a set of features, $x_1, x_2, ...x_k$, for a borrower i, the task is to predict a variable y_i, that is, the probability of default, $Pr(Default)$. The features x_i describes the borrower's credit history, for example, the number of lines of credit and the number of times the borrower defaulted in the past, among others. In practice, y_i is determined by long-standing credit scoring models, which are characterized by decisions such as whether the applicant had 90+ days delinquency in the two years after the opening of a new line of credit.

We frame the classification as a supervised learning problem in which we train

[4] https://aix360.mybluemix.net/data

a function f that can approximate the relationship $y_i = f(x_1, x_2..x_k)$. To train the function f, we use variety of ML models, including optimal classification trees, random forests, inductive logic programming, and neural networks. We describe these models next.

9.3.1 Random Forests

Random forests are an ensemble supervised learning technique. This technique aggregates multiple outputs from a set of predictors, in this case, multiple decision trees, in the belief that this will produce a more accurate classifier.

The key idea behind random forests is that a high-performing classifier can be constructed from a set of non-expert classifiers which are decision trees. A single decision tree is a supervised learning method that predicts the value of a target variable by learning simple if-then decision rules. It is constructed using the Classification And Regression Tree (CART) algorithm [40]. Each node in the decision tree is a condition on the value of a single feature that splits the data into two subsequent branches. CART recursively identifies the feature-value pair that best minimizes the tree's Gini impurity, a metric of the disorderliness of the labels of a set of data points.

Random forests are trained via a method called bootstrap aggregation, or bagging. The training data points are first randomly assigned into n groups with replacement, where n corresponds to the number of decision trees. Individual decision trees are fitted to a randomly chosen set of features in each group. To classify a new data point, the random forest aggregates the predictions from each of its constituent decision trees, and uses the majority vote as its classification. The random sampling of data points and features ensures that the resulting decision trees are uncorrelated. Thus, by aggregating their independent predictions, random forests are able to reduce variance and improve generalizability.

Random forests are difficult to interpret. For individual data points, each decision tree gives its if-else conditions that lead to a classification, but when these conditions are combined over the many trees of an ensemble, interpreting these conditions is impossible. This makes random forests effectively a black-box model.

9.3.2 Inductive Logic Programming

Inductive logic programming (ILP) [218] involves using first-order logic to represent and explain data. The dataset can be represented by a finite set of rules or clauses. In this section, the ILP developed is data-driven and clauses are learned from the data which is an input to the ILP model.

ILP requires that we specify the number of rules N, a given maximum rule size R, and the dimensional n binary input vector X. We construct Π, a $N \times R \times 2n$ tensor, and interpret softmax($\Pi[i, j]$) as a probability distribution for the jth term in the ith rule; that is, we obtain a $2n$-sized discrete probability distribution over the n features and their negations.

The rules are learned in a disjunctive normal form that consists of clauses with $\hat{\wedge}$ (AND) and $\hat{\vee}$ (OR) conditions. The AND and OR operators require binary operands. However, these operators must be extended to continuous operands to learn clauses from data. We use the product as a continuous extension of $\hat{\wedge}$, while a continuous extension of $\hat{\vee}$ is obtained from DeMorgan's Law, $\widehat{\vee}(x) = 1 - A(1 - x)$, where A is the continuous extension of $\hat{\wedge}$ (product as described earlier). From parametrization and using these continuous extensions of $\hat{\vee}$ and $\hat{\wedge}$ on $[0, 1]$, we can compute an approximated label $\hat{y}$ for an input vector X by concatenating X with its $1 - X$ to get a $2n$-vector X^*:

$$\hat{y} = \widehat{\bigvee_i}\widehat{\bigwedge_j}(X^* \cdot \text{softmax}(\Pi[i, j]))$$

where i ranges over the number of rules and j ranges over the size of a rule.

The binary cross-entropy loss between $\hat{y}$ and ground truth y can be minimized using stochastic gradient descent in order to learn the probability distributions in Π. Once the model is trained, we use the probability distribution Π to obtain a set of logical rules in disjunctive normal form, for all n from 1 to N and r from 1 to R, using the following expression:

$$\text{target} \leftarrow \bigvee_{n=1}^{N} \left(\bigwedge_{r=1}^{R} \text{argmax}_k(\Pi[n, r, k]) \right)$$

In our study, X is a numerical vector with each dimension corresponding to a different feature value. We first rank-normalize it to change its range to $[0, 1]$, and use the transformation T to binarize it: $T(x) = \sigma(a(X - b))$, where σ is the sigmoid function, a is the fixed scale parameter, and b is the parameter learned during optimization. We use stochastic gradient descent to learn the clauses based on the input data.

Using the inductive logic programming model, we can thus learn simple interpretable rules that can be used for classification and summarizing datasets.

9.3.3 Optimal Classification Trees

Decision trees are constructed in a top-down greedy way using the CART algorithm [40] as described in Section 9.3.1. At every node, the split is decided locally without the knowledge of other nodes. This makes decision trees only one-step optimal, leading to poor performance when classifying unseen points.

Optimal trees [30] are a variant of decision trees that are learned in a globally optimal manner. Optimal trees decide their split in one step, with knowledge of all other splits. The tree learning process is modeled as a mixed-integer optimization problem, which can be solved using fast available optimizers. Because optimal trees are constructed in a globally optimal fashion, they perform better than decision trees, and have all the advantages of other decision trees in terms of explainability. However, optimal trees become difficult to explain if they are very deep, and the rules learned at every node are complex. In addition, the number of variables involved in optimization for the creation of an optimal tree model is a linear function of dataset size and an exponential function of maximum depth. In general, mixed integer optimization does not scale well with a large number of variables. As a result, deeper optimal trees take a longer time to train on large datasets compared to decision trees, which limits their use for Big Data problems.

9.3.4 Neural Networks

Neural networks (NN) are supervised learning models loosely inspired by the biological networks of the human brain.

A NN is composed of three types of layers: an input layer, a number of hidden layers, and an output layer. An input layer relays the input features into the model, the hidden layers act as the computational engine, and the output layer generates the final model prediction. The input to each hidden unit is a linear combination of the units of the preceding layer. The hidden unit then computes its output by mapping its input through an activation function. A nonlinear activation function is commonly used to create nonlinear interactions between the units of the neural network. It is worth noting that logistic regression is a special case of a NN, with one hidden layer containing one hidden unit with a sigmoid activation function.

In NNs, the value of each hidden unit can be computed as:

$$h_j(x) = f(w_j + \sum_{i=0}^{n} w_{ij} \cdot x_i).$$

Here, w_{ij} is the weight from input x_i to hidden unit h_j. The weights w_{ij} can be learned by minimizing the loss function using optimizers like stochastic gradient descent.

The weights of the neural network create complex and nonlinear interactions between input features, which do not have human-interpretable meanings. However, the complex nonlinear interactions learned by NNs lead in general to better model performance. Moreover, the decision boundaries learned by NNs can be both high-dimensional and extremely nonlinear. Thus, learned features may have varying significance at different points of the feature space.

9.4 Data

For our task, we use a home equity line of credit (HELOC) dataset provided by FICO. A HELOC is a revolving loan in which the collateral is the borrower's equity in their house. Like a credit card, a HELOC is available for a set time frame during which a borrower can withdraw money as needed. This dataset was released as part of the explainable machine learning challenge by FICO[5], and is widely used for explainable ML studies (see also Section 9.2). Therefore, we follow this literature and use the same dataset to study the explainability of several ML models in Section 9.3. However, our methodology can easily be generalized for other credit risk datasets.

The FICO dataset contains 10,459 borrowers who were granted HELOCs during a two-year application window from March 2000 to March 2002. In March 2003, a year after the application window had closed, a performance snapshot was captured, and the risks of the borrowers were evaluated. In this dataset, an applicant might have used their HELOC for a duration between one and three years, depending on the time of their approval. The dataset contains 5,459 non-creditworthy records, while the remaining 5,000 records are deemed creditworthy. A record is considered non-creditworthy if the consumer has been 90 days past due at least once over a period of 24 months since opening the credit. The record is considered creditworthy if the consumer has been making payments without ever being 90 days delinquent.[6] In addition to the binary target variable of credit risk classification, each credit applicant is characterized by 23 predictor features, 21 continuous and 2 categorical. Further information about these features, and important terms relevant to the dataset, are provided in Appendix G.1.2.

Although our dataset is relatively small compared to the total HELOC market, it is already sufficient to provide a proof-of-concept benchmark that it is possible to improve credit predictions using machine learning models and, more importantly, to explain these machine learning models using our proposed framework. With a much bigger and more complete dataset in practice, one can expect to explain these models in a more precise and granular fashion.

[5] `https://community.fico.com/s/explainable-machine-learning-challenge`

[6] In general, default is a rare event. However, in this public dataset, the classification is provided by FICO and the target variable is relatively balanced between the two classes. Though this is a re-sampled dataset compared to the population of all HELOCs, it is still widely used in the literature especially when studying model interpretability (see, for example, [244], [59], and https://aix360.mybluemix.net/data.)

9.5 Evaluation

In this section, we evaluate the different models described in Section 9.3. These machine learning models, including optimal trees, random forests, and neural networks, are trained by recoding all the features using the weight of evidence encoding. An inductive logic programming model is trained using the methodology described in section 9.3.2. We also evaluate the neural network model trained on a one-hot encoding version of the dataset. The details of these implementations and best-performing models are included in Appendix G.2 for reproducibility.

We include the two-layer additive risk model in our evaluation [59], which was the best performing model of the FICO Data Challenge[7]. It partitions features into different subgroups, combining scores from different subgroups into a global model score. Its feature subgroups are generally interpretable, as they are created through the intervention of an expert in the field. The model resembles a two-layer sparse neural network. This model serves as a baseline for the other black-box models.

We evaluate these models using K-fold cross-validation. In K-fold cross-validation, the input dataset is randomly partitioned into K equal subsets. In each run, one of the subsets is chosen as the test set and the remaining as the training set. For this analysis, we choose $K = 5$, i.e., in each run of cross-validation, the training set contains 80% of the dataset ($7,898$ points), and the test set contains the remaining 20% ($1,973$ points). The class distributions of the train-test sets are included in Table G.1 in the appendix. The standard errors of the performance measures are obtained by bootstrap sampling of the dataset (keeping train/test split of 80-20%) followed by re-training of the model and estimating the performance metric for 1000 iterations. The performance measures for 1000 iterations can be used to obtain standard errors.

We evaluate the models using different metrics: accuracy, area under the curve (AUC), false positive rate (Type I error) and false negative rate (Type II error). The metrics are averaged across all five cross-validation datasets. The AUC measures the ability of a classifier to distinguish between the classes. Our dataset is fairly balanced,

[7]http://dukedatasciencefico.cs.duke.edu/

as is illustrated in Table G.1, hence the accuracy is also a good measure of goodness of fit along with the AUC. All models predict the $Pr(Default)$ values for every sample. To classify the different samples, we used a threshold of 0.5.

As Table 9.1 illustrates, we find that the NN model trained on weight-of-evidence (WoE) data[8] has the best performance, with an accuracy of 74.75%. It outperforms the NN trained on one-hot encoded data, implying that the WoE-encoded features have more information than one-hot encoded features. The two-layer additive risk model performs second best, with an accuracy of 74.12%. This model is easy to interpret; however, it requires the input and involvement of experts who know about the constraints in the dataset and the relationship between its different features (in order to divide features into different subgroups) that might not always be available for various applications.

On the other hand, the rule-based models, optimal trees (black box) and random forests suffer from the problem of explainability. For random forests, analyzing the ensemble of 140 trees at a time is intractable. Similarly, the optimal trees (black box) model has a single deep tree with multiple features deciding every split, which is difficult to interpret. If we train a shallow optimal tree, its accuracy drops to 72.28%, but the model is easy to interpret and analyze. We also obtain a small set (1 or 2) of simple rules from inductive logic programming that are easy to understand, but they come at the cost of a decrease in accuracy. Even though interpretable rule-based models do not perform as well as NN model, we will see in Section 9.6 that they are useful for explainability for the different stakeholders involved.

Using ILP, we obtain a single simple rule, that is, “if $ExternalRiskEstimate$ is smaller than 72, then classify borrowers as non-creditworthy,” achieves an accuracy of 70.71%. $ExternalRiskEstimate$ is a condensed version of the borrower’s credit risk, a metric similar to the FICO score.[9] In general, companies use the credit score and a threshold associated with it to decide an individual’s creditworthiness; in our case,

[8] Appendix G.1.4 provides a summary of the WoE feature pre-processing

[9] Although this may not be an interpretable feature by itself, we keep this feature in the analysis because it is widely used and our goal is to generate an explanation methodology given a dataset and model irrespective of features included in the study.

Models	Test Accuracy	Test AUC	False Positive Rate	False Negative Rate	Precision
ILP (1 rule)	70.71(0.92)	70.64(0.93)	31.13(1.38)	27.57(1.20)	71.56(1.24)
ILP (2 rules)	70.92(0.94)	70.62(0.94)	36.98(1.48)	**21.75**(1.21)	69.59(1.21)
Optimal Trees (Interpretable)	72.28(1.03)	74.18(1.74)	28.25(4.26)	27.21(2.99)	71.59(2.16)
Optimal Trees (Black Box)	74.12(0.46)	74.46(0.50)	29.16(2.63)	22.85(2.78)	73.54(1.14)
Random Forest (140 Trees)	73.77(0.01)	79.82(0.01)	29.71(0.02)	23.01(0.02)	73.61(0.02)
Two Layer Additive Risk	74.12(0.88)	81.01(0.83)	30.31(1.28)	21.77(1.17)	73.63(1.15)
NN (one-hot)	74.10(0.88)	80.59(0.84)	**25.16**(2.50)	26.60(2.25)	74.06(1.53)
NN (WoE)	**74.75**(0.98)	**81.40**(0.84)	28.31(1.56)	22.42(1.37)	**74.70**(1.30)

Table 9.1: 5-Fold cross-validation performance of ML models.
The NN trained on WoE data performs the best with a mean test accuracy of 74.7% and AUC of 81.4%. A higher AUC implies the NN model is better able to distinguish between the creditworthy and non-creditworthy classes compared to other models. All rule-based classifiers have smaller values of AUC compared to the others. The NN (one-hot) has the minimum false positive rate, while the two-layer additive risk model has the minimum false negative rate when the threshold used to classify the model's output is 0.5. All the reported values are in percentages. The values in the brackets "()" are the standard errors of the estimates.

it is the *ExternalRiskEstimate* with a threshold of 72. We observe that the use of machine learning models using different features about a borrower's credit history can increase the accuracy of the delinquency forecast by four percentage points compared to the naïve use of credit scores.

Credit companies are naturally interested in using models with the best performance. If credit companies give loans to borrowers by misclassifying defaulters as non-defaulters (that is, false negatives), they will suffer losses. Similarly, if credit companies deny loans to borrowers who can repay (that is, false positives), they will lose business. A four-percent improvement in accuracy may not seem big, but given the enormous size of the mortgage business in the United States, even a small increase in model performance can create a substantial economic impact.[10] Moreover, a more accurate model can lead to more borrowers having access to loan opportunities in aggregate.[11] However, regulatory practice and the black-box nature of the models prevent them from harnessing these benefits. This motivates us to analyze the explainability of these different models.

9.6 Explaining Machine Learning Models

We have shown in Section 9.5 that the neural network model outperforms every other model tested. However, the non-interpretable nature of neural networks limits its adoption. In this section, we analyze the explainability of models in order to help their wider adoption for different applications.

The multiple parties involved in credit risk management require different explanations for different purposes. For instance, loan applicants who are denied loans are interested in finding the reason for the denials and suggestions that can make them more creditworthy. Data scientists, on the other hand, are more interested in understanding the data, while regulators demand fairness from the models and analyze the model's behavior in extreme scenarios. We characterize the kinds of explanation

[10] In practice, with more data available by real lenders, the improvement in model performance should be even bigger than our proof-of-concept study.

[11] See, for example, the model of [117].

that are appropriate for the following end users: loan companies, data scientists, loan applicants, and regulators.

9.6.1 Interpretability for Loan Companies: Opening the Black Box

Loan companies use machine learning models to evaluate the creditworthiness of a borrower. Regulations require the loan companies to give a set of reasons for every denial of the application. Consequently, loan companies are interested in finding the factors that contribute to the creditworthiness of the individual. In addition, they require explanations for every prediction of the model. These explanations make the model transparent, and they assist in finding the most representative samples (borrowers from the past) for a data point (a new borrower). We use state-of-the-art methods to generate explanations for our model predictions, including LIME [239] and SHAP [195]. We generate explanations for the best-performing NN model. Our key contributions in this section include applying and comparing existing models (LIME and SHAP) to generate explanations for loan companies, modifying LIME to improve the validity of linear approximations in our context, and using the k nearest neighbor (kNN) algorithm on the explanation to find representative samples for a data point.

LIME

LIME is a model-agnostic technique that approximates the decision boundary of a model at a particular data point using a linear approximation[12]. This linearity makes the LIME model interpretable. The approximation is constructed by training a locally-weighted linear regression model in the neighborhood of the data point of interest. The coefficients of the regression can be used to justify the data point's classification. Because the features are on the same scale (weight-of-evidence encoded), the coefficients are comparable.

[12] The key idea behind LIME is that every segment of the decision boundary begins to look linear at increasingly smaller scales.

Constructing a locally-weighted linear regression model involves sampling and perturbing the data points that are used for training. One of the shortcomings of LIME is that the perturbed data points sampled by LIME may be invalid. For example, imagine a dataset with two features A and B, with a constraint that A < B. Sampling each feature's value independently, as is done in LIME's original algorithm, may produce perturbed data points where this constraint is violated. We address this shortcoming of LIME by modifying the algorithm to take into account the potential interdependencies of different input features.

This modification changes the methodology of sampling a data point of the LIME algorithm. In its original implementation, a data point with n features is sampled by independently sampling each of its n features from univariate normal distributions. In our modified implementation, a data point is sampled directly from a joint multivariate normal distribution across all features. This allows perturbations to be informed by the correlation of features. The multivariate normal distribution is still centered on the mean of each feature value, but the standard deviation along each feature is determined by the correlation matrix of the input data.

We first evaluate our modified LIME with respect to the validity of the perturbed data points. As noted, LIME samples perturbed points in the neighborhood of the input data point. We generate a set of $5,000$ perturbed data points from several univariate normal distributions centered on the feature means and a single multivariate normal distribution centered similarly. We measure the quality of the sampled points by two metrics.

Correlation. The correlation between features of the perturbed data points should be similar to that of the features in the training dataset. We compute the mean-squared error (MSE) between the correlation matrix of the training data and that of the perturbed points generated by both implementations of LIME. Let us call these values $MSE^{original}$ and $MSE^{modified}$, respectively. We find that $MSE^{modified}$ is much smaller than $MSE^{original}$ (0.000 vs 0.053). This is expected because the correlation matrix of the training data is an input to the multivariate normal distribution that samples perturbed points in our modified implementation. Our perturbed data more closely resemble the characteristics of the original dataset.

Constraint Violation. There are 12 constraints relevant to the HELOC dataset (6 relational constraints and 6 value constraints). We determined these via discussions with FICO personnel. Examples of these constraints include, "all feature values with percentage units should be smaller than 100," and "the number of lines of credit not delinquent must be less than the total number of lines of credits." The 5,000 perturbed data points are scanned for violations of these constraints. The results for all constraints are shown in Table 9.2. We see that the modified implementation of LIME produces fewer violations than the original implementation in 7 out of the 12 constraints. Moreover, a perturbed point sampled by the modified implementation has on average 1.954 constraint violations, versus 2.627 for the original implementation. Hence, the modified LIME does provide an improvement in terms of generating more realistic data points.

We also compute the average goodness of fit of the linear model in LIME for all test data points. We find that the average R^2 of the modified LIME is 90%, compared to 88% for the original LIME. Comparing correlation, constraint violation, and goodness of fit, we conclude that the modified LIME leads to better surrogate models. In the remainder of the paper, we use this modified LIME for our analysis, which we refer to as the LIME model for simplicity.

Explaining Model Predictions (LIME)

LIME learns a linear surrogate model. As a result, for each data point, the coefficients of its linear regression can be interpreted as the change in the output produced by a unit change in the corresponding feature value, given that other feature values are held constant. Figure 9-1 shows an example of a LIME explanation. A unit increase in the values of features with coefficients shown in red lines will increase the non-creditworthy probability, while a unit increase in the values of features with coefficients shown in blue lines will decrease the non-creditworthy probability.

Index	Constraint	Original LIME Violations	Modified LIME Violations
1	All feature values interpreted quantitatively must be non-negative	4383	4024
2	PercentLOCNeverDelq $\leq$ 100	1198	1183
3	PercentInstLOC $\leq$ 100	1	2
4	PercentLOCWBalance $\leq$ 100	340	296
5	FracRevLOCLimitUse $\leq$ 100	62	75
6	FracInstLOCUse $\leq$ 100	509	506
7	NumLOC90PlusDaysDelq $\leq$ NumLOC60PlusDaysDelq	1656	655
8	NumLOCReqLast6MExPastWeek $\leq$ NumLOCReqLast6M	2095	388
9	NumLOC60PlusDaysDelq $\leq$ NumTotalLOC	191	239
10	NumLOC90PlusDaysDelq $\leq$ NumTotalLOC	192	233
11	NumLOCNotDelq $\leq$ NumTotalLOC	2257	1915
12	NumLOCInLast12M $\leq$ NumTotalLOC	249	255

Table 9.2: Constraints Violation
Comparison of constraint violations among 5,000 perturbed points sampled from the original and modified implementations of LIME. The points sampled from the modified implementation of LIME have fewer violations in 7 out of the 12 constraints.

After obtaining the feature importance for all the data points, they can be aggregated to find the overall feature importance. In Figure 9-2, we look at the global feature importance for the model. The top three most significant features are the months since the newest request for a new line of credit (excluding those requested in the past week), the external risk estimate, and the fraction of all revolving line of credit limits in use. It is worth emphasizing that the external risk estimate is an important feature of the model, but not the most important one. This can be understood by observing the nonlinear nature of the classifier in Figure 9-2a. For example, for the feature "months since the newest request for a new line of credit (excluding those requested in the past week)," having small and large values both contribute to a decrease in the non-creditworthy probability, while having values around the median contributes to an increase in the non-creditworthy probability. Such a nonlinear relationship would be difficult to capture using traditional models. These generated

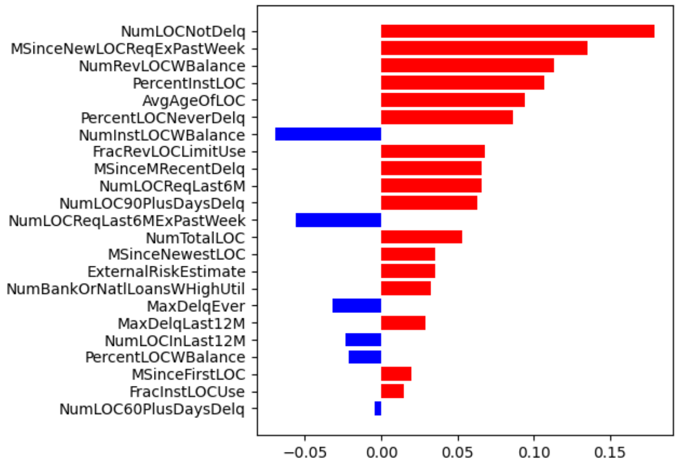
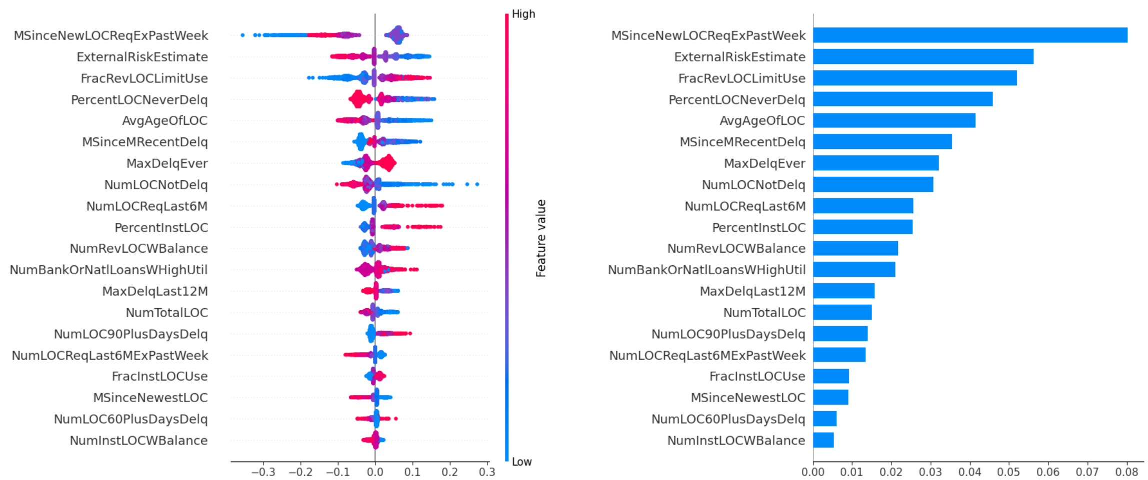


Figure 9-1: LIME Approximation
The LIME model approximation for a random data point. A unit increase in features with a red horizontal line will increase the non-creditworthy probability. Similarly, a unit increase in features with a blue horizontal line will decrease the non-creditworthy probability. The contribution of each feature in the model explanation can be obtained by multiplying each feature coefficient by the feature value.

explanations and the overall feature importance also aid in discovering biases in the model.

The quality of LIME explanations depends on the fidelity of the LIME model. We evaluate the fidelity of the approximations produced by our modified implementation of LIME for the best-performing NN. We do this by constructing an approximation at each of the 1973 points in the test dataset and computing the fraction of times the NN and the LIME approximations produce the same classification. We find that out of 1973 data points, 1935 points have the same classification for the linear LIME approximation and the NN model being analyzed. The high fidelity of the LIME approximations shows that LIME is able to generate valid linear approximations that can be trusted for a large number of data points.

However, for the few cases where the classification of the NN model and LIME do not match and are different significantly, the linear approximation cannot be trusted. In order to generate explanations for these points, we discuss another method.



(a) Feature contribution for all sample in test set

(b) Average feature contribution (absolute value) for all samples

Figure 9-2: Feature Importance LIME

Feature importance obtained by aggregating LIME explanations for all samples in the test set. In Figure (a), the color coding (blue-red) represents the value of the feature. For example, for the external risk estimate, the larger values (in red) contribute to the decreasing non-creditworthy probability, while the smaller values (in blue) contribute to the increasing non-creditworthy probability. The contributions are defined relative to 0.5 (equal probability of default/non-default). The x-axis is the contribution of the individual feature in the output of the model relative to 0.5. For example, for a data point, if the contribution of all other features is zero and the contribution of external risk estimate is -0.2, the model's output will be $0.5 - 0.2 = 0.3$. Figure (b) is obtained by aggregating the feature importance as obtained for all points, which illustrates that the three foremost significant features are months since the newest request for a new line of credit (excluding those requested in the past week), the external risk estimate, and the fraction of all revolving line of credit limits in use.

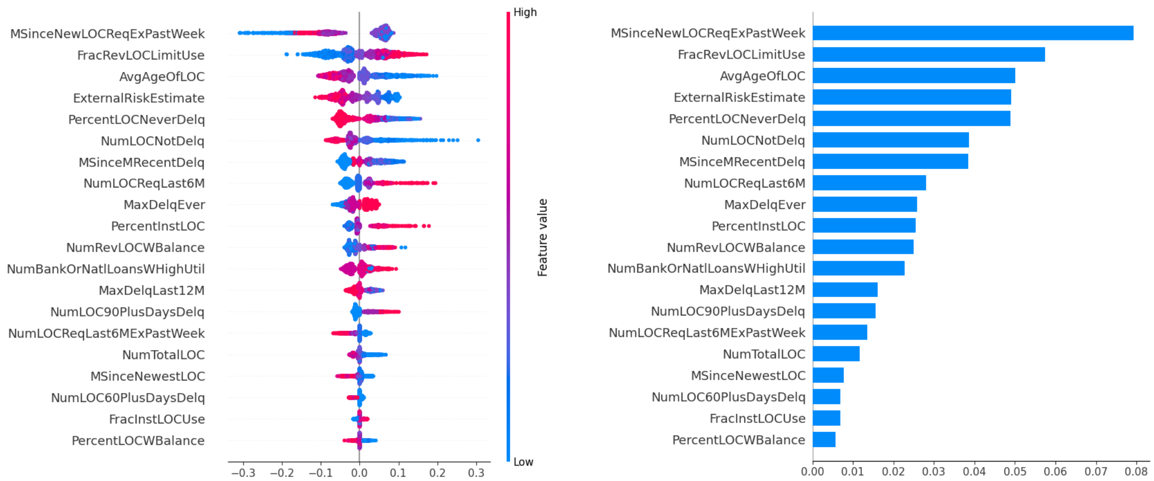
SHapley Additive exPlanations (SHAP)

SHAP is an explainable AI method with an economic foundation. It performs the Shapley value decomposition of the model output, giving the contributions of every feature at a data point. Like LIME, it is also a model-agnostic method.

Shapley values have several good properties that satisfy a number of important criteria, including local accuracy (i.e., the model explanation matches the original prediction), handling of missing data (i.e., if the feature is absent, its contribution to the model prediction will be zero), and consistency (i.e., if the model changes in a way that leads to larger marginal contributions for a feature, the Shapley values also increase). Shapley values also incorporate the interactions between features in the process of calculating feature importance, making SHAP a more reliable method for interpretability than LIME.

We use the Kernel SHAP implementation from [195]. The Kernel SHAP procedure computes the Shapley values by running a weighted-least-squares regression whose solution is the Shapley values of features. To explain a point x_i, the different points used in the linear regression are obtained by selecting a subset of features from x_i, and the remaining subset of features not selected are replaced with values from background data points (that is, from training data). The weights of the linear regression are decided by the size of the sampled subset. For example, a subset of one feature has the maximum weight because it provides the maximum information about that feature's contribution. Using the described weighted-least-squares regression, we obtain Shapley values for the data point x_i. It is worth highlighting that the Kernel SHAP implementation relies on taking subsets of features, that is, $2^{Numberoffeatures}$. As a result, Kernel SHAP does not scale well.

Using SHAP, we obtain explanations for the classification of all the data points in the test set. Similar to LIME, the Shapley values for features can be aggregated for all test points to find the overall impact of a feature on the model. We present the feature importance for all data points in the test sample in Figure 9-3, which shows that the months since the newest request for a new line of credit (excluding those requested in the past week), the fraction of all revolving lines of credit limits in use, and the average age of lines of credit are the top three most important features for the model. We observe a minor difference in the feature contribution obtained from the LIME and SHAP. This may be due to the fact that not all LIME approximations are locally accurate.



(a) Feature contribution for all samples in the test set.

(b) Average feature contribution (absolute value) for all samples.

Figure 9-3: Feature Importance SHAP
Feature importance (Shapley values) obtained by aggregating Kernel SHAP explanations for all the samples in the test set. In Figure (a), the color coding (blue-red) represents the value of the feature. For instance, the larger values (in red) for the external risk estimate contribute to the decreasing non-creditworthy probability, while the smaller values (in blue) contribute to the increasing non-creditworthy probability. The contributions are defined relative to 0.5 (equal probability of default/non-default). The x-axis is the contribution of the individual feature in the output of the model relative to 0.5. For example, for a data point, if the contribution of all other features is zero and the contribution of external risk estimate is -0.2, the model's output will be $0.5-0.2 = 0.3$. We obtain the overall feature importance (as illustrated in Figure (b)) by aggregating Shapley values for all test points. The most important features are the months since the newest request for a new line of credit (excluding those requested in the past week), the fraction of all revolving lines of credit limits in use, and the average age of lines of credit. These are largely consistent with the LIME feature importance. Small discrepancies can be attributed to the inaccurate local approximations of LIME.

It is also worth emphasizing that both LIME and SHAP have their weaknesses, which may affect the explanations in the credit risk domain. For example, in addition to feature importance, we can obtain locally linear approximations using LIME (as obtained in Figure 9-1), which are not available in SHAP. These local approximations of the decision boundary are necessary for other applications, as will become clear in the next section. On the other hand, the explanations using LIME might not satisfy important properties like local accuracy and consistency that Shapley values allow. In addition, we have incorporated feature correlations in the sampling of data points for learning the local linear model in LIME, which leads to fewer constraint violations in the input feature vector, hence creating more realistic data points. However, in SHAP such feature correlations are not considered when computing Shapley values.

From the computational perspective, different runs of LIME produce slightly different explanations because of the randomness in sampling local data points. Consequently, it can potentially suffer from instability issues. On the other hand, computing Shapley values is not scalable for datasets with a large number of features. In such scenarios, LIME approximations can be computed efficiently. Therefore, both methods (SHAP and LIME) should be used in practice depending on the use case, carefully and judiciously.

Finding Representative Samples

Another application of model explanations is to find the most representative data point for a particular target point analyzed. It can be useful in scenarios where loan companies might require representative data points from the past to explain the model prediction of a particular loan candidate. For example, the borrower x_i might default in the future because it is similar to x_j and x_k who defaulted in past.

The k-nearest neighbors algorithm (kNN) applied to the data naively can generate the most similar points in the sample. However, if we apply kNN to the original feature space, the less significant features will end up contributing to the distance calculations between the data points. To avoid this problem, every data point can be represented by a feature importance vector (for instance, a vector of Shapley values), and kNN can be applied to the feature importance vector. Using SHAP will ensure that less important features have smaller feature contributions (that is, smaller Shapley values). Hence, the contributions of less important features to the distance calculation in the kNN will be minimal. This modification leads to finding a more representative data point for each new data point being analyzed.

9.6.2 Interpretability for Regulators: Model Fairness and Stress Testing

In the previous subsection, we generated explanations for model predictions. Apart from concerns about the transparency of the credit approval algorithm, government regulations mandate that these algorithms must be fair. Concerns about the fairness of a model arise when the model gives importance to features like race, religion, or gender that are specified in law. Our HELOC dataset does not contain any such features overtly. Nevertheless, the fairness of a black box model can be examined (if features were present) using the individual and overall feature contributions that are obtained from LIME and SHAP.

Another important aspect of interpretability for regulators is stress-testing the models. The credit risk models used by companies need to withstand appropriate stress testing. This involves simulating extreme scenarios and analyzing the model's behavior in response to generally extreme macroeconomic conditions.

Extreme macroeconomic conditions cannot be simulated in our model due to its lack of macroeconomic features. However, we demonstrate how interpretability methods can help in stress testing a black-box model for extreme customers. The same methodology can be generalized for extreme macroeconomic conditions. We use two extreme customers in our stress testing. In the first case, an extreme customer is obtained by sampling points from a multivariate normal distribution with extreme values. While sampling, some feature values may no longer satisfy the feature-specific constraints; for them, we manually truncated the feature values. In the second case, we use a customer whose feature values are all -9, that is, there is no bureau record or investigation on file about this customer. We have multiple such data points in the dataset that are not included for training or testing purposes. The extreme points used in our analysis are included in Table 9.3.

Feature	Data Pt 1	Data Pt 2
ExternalRiskEstimate	131	-9
MSinceFirstLOC	935	-9
MSinceNewestLOC	99	-9
AvgAgeOfLOC	467	-9
NumLOCNotDelq	89	-9
NumLOC60PlusDaysDelq	6	-9
NumLOC90PlusDaysDelq	5	-9
PercentLOCNeverDelq	100	-9
MSinceMRecentDelq	115	-9
MaxDelqLast12M	16	-9
MaxDelqEver	13	-9
NumTotalLOC	125	-9
NumLOCInLast12M	6	-9
PercentInstLOC	88	-9
MSinceNewLOCReqExPastWeek	35	-9
NumLOCReqLast6M	13	-9
NumLOCReqLast6MExPastWeek	12	-9
FracRevLOCLimitUse	71	-9
FracInstLOCUse	100	-9
NumRevLOCWBalance	23	-9
NumInstLOCWBalance	25	-9
NumBankOrNatlLoansWHighUtil	21	-9
PercentLOCWBalance	100	-9
Model Prediction	0.53	0.75

Table 9.3: Extreme Customers (data points).
The first point is sampled, assuming features are multivariate Gaussian. The second point is obtained by setting all feature values to -9 (no bureau record or investigation). Since the data point 1 is sampled from a multivariate Gaussian distribution, it may contain feature values that are not coherent.

Regulators are interested in verifying the validity of the model's output and analyzing its behavior. We preprocessed features for these extreme customers using the weight of evidence encoding, which leads to extreme values being assigned to one of the bins. Due to this binning, the model was able to produce valid predictions. For case 1, the output is $Pr(default) = 0.53$, which implies that the person may or may not default with almost equal probability. Since this data point is sampled randomly, we cannot comment on the accuracy of the prediction. For case 2, the model's prediction is $Pr(default) = 0.75$, which implies that, when there is no information about the person on file, the person is likely to default, hence preventing lenders from approving credit for such individuals with no information in their credit file.

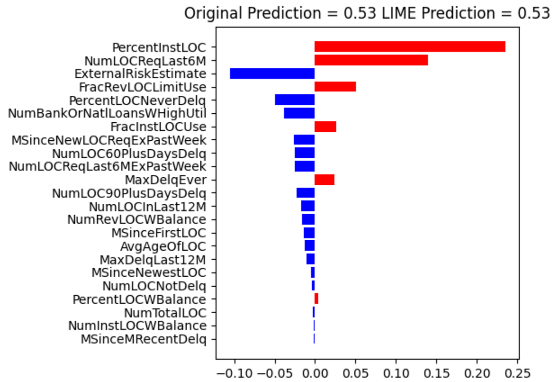
To understand the behavior of the model in the proximity of extreme points, we use the LIME model approximation. For cases 1 and 2, we present the LIME approximation in Figure 9-4. In both cases, the predictions from the LIME model are close to the original model predictions. For extreme customer 1, the model's output is 0.53 and the LIME prediction is 0.53. For extreme customer 2, the model's output is 0.75 and the LIME prediction is 0.68. We also compute the R^2 for the LIME linear model, which is 99.6% and 88.8% respectively for the two customers. These results demonstrate that the approximations are fairly trustworthy, and may be useful to regulators to learn more about the model's behavior in extreme scenarios.

In this section, our key contributions include a proof-of-concept demonstration of the use of explainable AI methods for generating explanations in model fairness testing and stress testing that may be able to assist regulators in their duties. Next, we discuss the use of model interpretability for loan applicants.

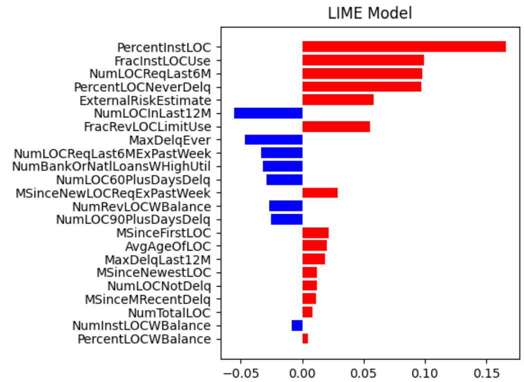
9.6.3 Interpretability for Loan Applicants: Counterfactual Suggestions

Loan applicants are interested in two major aspects of model interpretability. The first is to learn the reason behind their denial or approval, and the second is how to modify the features that might change their classification in the future. Explanations for the reason behind denial or approval can be obtained from the techniques discussed in the previous sections. In this section, we discuss the methodology of generating instructions for reversing the model's classification using state-of-art models in the literature and make appropriate modifications to the existing models to generate realistic instructions for reverse classification.

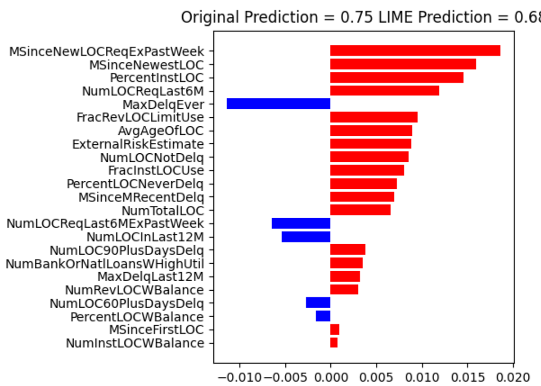
Applicants are interested in counterfactuals that provide information about the



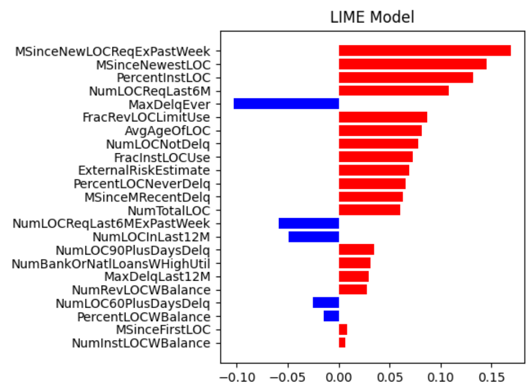
(a) Case 1: Feature Contribution.



(b) Case 1: LIME Model



(c) Case 2: Feature Contribution



(d) Case 2: LIME Model

Figure 9-4: Extreme Scenarios: LIME

LIME explanations for the two extreme scenarios (case 1: an extreme sample, case 2: no information in file). The regulators obtain the linear approximation using the LIME model in the proximity of an extreme sample. In both cases, the LIME approximation is close to the model's output. For extreme customer 1, the model's output is 0.53 and the LIME prediction is 0.53. For extreme customer 2, the model's output is 0.75 and the LIME prediction is 0.68. The R^2 of LIME is 99.6% and 88.8%, respectively. The LIME models are $\Sigma(FeatureValue) * (LIME - Coefficient)$ whose coefficients are represented by the horizontal lines. We can observe that for extreme case 1, the model gives its maximum importance to the percentage of lines of credit that are installment lines of credit, while for extreme case 2, the model gives its highest importance to the months since the newest request for a new line of credit excluding those requested in the past week. The above analysis can be extended to different extreme conditions.

steps that might change the decision of the model. Consequently, the counterfactuals being generated should have the following properties: reverse classification (that is, the model prediction for the counterfactual should be reversed from the original decision), proximity (that is, the counterfactual should be close to the original data point), and diversity (that is, there should be multiple different counterfactuals from which an individual be able to choose).

We generate counterfactuals using the state-of-the-art method, Diverse Counterfactual Explanations (DiCE) [216]. This method learns the counterfactuals c_i for a data point y optimizing over the requirements of the counterfactual. In particular, it learns the c_i's by minimizing the following loss:

$$L = \sum_{i=1}^{k} loss(f(c_i), y) + \lambda_1 \sum_{i=1}^{k} dist(c_i, y) - \lambda_2 * diversity(c_1, c_2 ... c_k)$$

where the first part is the reverse classification loss (cross entropy for our application), the second part is the proximity to the original data point (average $l1$ distance), and the third part is the diversity component (determinant of kernel matrix given by the distance between counterfactuals). λ_1 and λ_2 are the weights for proximity and diversity components, respectively. The exact details of the implementation of different loss function components can be found in [216].

We study a few illustrative examples and analyze the properties of the counterfactuals generated using DiCE.[13] Counterfactuals can be generated using other algorithms as demonstrated in [129]. We use DiCE because it implements the diversity functionality while generating counterfactuals. We analyze the importance of diversity next.

Applicants classified as non-creditworthy are interested in the steps that can make them creditworthy. The steps suggested must be practical enough for an applicant to implement. Some features are inherently impossible for individuals who are deemed non-creditworthy to improve. For example, the maximum delinquency ever in days

[13] We use the publicly available implementation of the DiCE algorithm, found at `https://github.com/interpretml/DiCE`.

cannot be decreased: it is an event that happened in the past, and cannot be changed. Likewise, the total number of lines of credit established cannot be increased, since the person has been turned down from opening a new line of credit.

In our system, the features that are impossible to modify are incorporated as constraints in the optimization of loss L. Table 9.4 shows an example of the set of suggestions to become creditworthy for an individual who is deemed non-creditworthy. All of the suggestions are possible to implement, though some might be difficult. In cases where it is hard for an individual to implement certain changes, he or she may be able to choose from a diverse set of counterfactuals to reverse the classification.

In Table 9.4, the first counterfactual suggests increasing the months since the applicant's most recent delinquency to more than four years. However, this requires an individual to wait for four years while making payments for all his lines of credit. In this case, the individual may instead choose to follow another set of steps, as suggested by the third counterfactual: increasing the percentage of lines of credit never delinquent (by decreasing the total lines of credit), along with decreasing the number of lines of credit requests made in the six months prior to applying for a new line of credit. Similarly, diversity-constrained counterfactual suggestions can be made for all individuals who are deemed non-creditworthy, and individuals can follow the steps which are most convenient to their circumstances.

Likewise, the people who are classified as creditworthy would like to maintain their creditworthy status. Hence, they would like to know which actions to avoid that might make them non-creditworthy. Table 9.5 presents one such example. The applicant in Table 9.5 might face a credit denial if he or she takes steps such as opening up new lines of credit, decreasing the lines of credit that are not currently delinquent, increasing the fraction of revolving lines of credit, and making new requests for line of credit in the six months prior to applying for a line of credit. These steps result in the non-creditworthy probability as suggested by the model increasing from 0.09 to 0.58, which may lead to a denial of credit. From these counterfactuals, the applicant thus knows the steps not to take that may decrease their creditworthiness.

To gauge the ability of DiCE to generate counterfactuals, we run it for all test data

Feature	Original Value	New Value
MSinceMRecentDelq	$0-4$	$48-$
NumLOCReqLast6M	$2-4$	$1-2$
$Pr(Default)$	0.63	0.44
AvgAgeOfLOC	$75-98$	$98-$
NumTotalLOC	$9-14$	$0-1$
NumLOCReqLast6M	$2-4$	$1-2$
$Pr(Default)$	0.63	0.43
PercentLOCNeverDelq	$0-82$	$98-$
NumLOCReqLast6M	$2-4$	$1-2$
Pr(Default)	0.63	0.31

Table 9.4: Counterfactual
The steps toward creditworthiness suggested to an individual deemed non-creditworthy. Following these steps, an individual should be able to decrease the non-creditworthy probability predicted by the machine learning model. We present three diverse counterfactuals. The first counterfactual suggests changing the months since the most recent delinquency from the value of 0-4 months to greater than 48 months, which means that person should not be delinquent for the next four years. In the second counterfactual, one of the suggestions is to decrease the total number of lines of credit to 0-1 (by closing accounts). Because some suggestions may be difficult for an individual to follow, we provide multiple diverse counterfactuals. The individual can follow the third counterfactual, which suggests decreasing the number of lines of credit to increase the percentage of lines of credit never delinquent, and decrease the number of lines of credit requests in the most recent six months prior to applying for a new line of credit. Providing multiple diverse counterfactuals allows the option of selecting the most convenient option for an individual.

points (total: 1973) and count the number of points for which the counterfactuals were successfully found. In Table 9.6, we show that increasing the proximity constraint leads to a decrease in the average number of feature changes in the counterfactuals obtained. We also observe that increasing the proximity constraint leads to a decrease in the number of successfully generated counterfactuals. An optimal algorithm to generate counterfactuals for an individual thus is to start with a high value for the proximity constraint and slowly decrease it until a counterfactual is found.

Using DiCE and related counterfactual generating methods, we can generate suggestions for loan applicants, as discussed above. One limitation of the counterfactual generation algorithm is the inability to take feature relationships into consideration.

Feature	Original Value	New Value
AvgAgeOfLOC	$60-75$	$4-29$
NumLOCNotDelq	$12-17$	$4-6$
NumLOCReqLast6M	$0-1$	$1-2$
FracRevLOCLimitUse	$13-29$	$77-$
$Pr(Default)$	0.09	0.58

Table 9.5: Counterfactual Creditworthy
The steps that can make a creditworthy-deemed individual non-creditworthy, as suggested by the algorithm. In simple English, if the individual decreases the number of lines of credit not currently delinquent, opens up multiple new lines of credit leading to a decrease in the average age of their lines of credit, made multiple requests for lines of credit in the past six months, and increases the fraction of their revolving line of credit, then the individual will be deemed non-creditworthy.

Proximity Constraint(λ_1)	Mean # of Feature Changes	Loss Value (Distance)	# of success counterfactual found
0.5	3.96	0.054	1961
1.5	3.35	0.045	1909
5.0	2.66	0.035	1854

Table 9.6: Proximity Analysis.
By increasing the proximity constraint (λ_1), the number of changes required for generating a successful counterfactual decreases, and the number of data points for which successful counterfactuals are generated also decreases. An optimal algorithm to generate counterfactuals for an individual is to start with a larger proximity constraint and slowly decrease it until a counterfactual is found.

For example, if one of the generated counterfactuals suggests decreasing the number of the total lines of credit, that also leads to modifications in the features that depend on the total number of lines of credit. However, this type of relationship is ignored in the present method. Solving this problem would involve the intervention of an expert who is familiar with the relationship between the model's features, and can encode them as constraints in the counterfactual generation optimization process. We leave this for future studies.

9.6.4 Interpretability for Researchers and Data Scientists: Simple Rules to Summarize the Dataset

One aspect of interpretability of particular interest to data scientists is the summarization of the data and the model, and answering key interpretability questions for other stakeholders. A global view of the data and the model is able to give the researcher an idea about any possible problems with the model. It can additionally help them to present a summary of the model to managers or regulators. In this section, we discuss methods that may help data scientists to demystify these black box models for their particular needs. The key contribution in this section includes applying data-driven inductive logic programming and decision trees to generate a dataset summary which is not discussed in earlier sections.

We have already discussed the first aspect of interpretability, the summarization of the model. A good summary of the model includes determining the most important features that contribute to the model's prediction. Using the methods described in the previous section (LIME and SHAP), data scientists can obtain the most important features and know their contribution to the model's overall prediction. For example, Figure 9-3 illustrates that the feature, "months since the newest request for a new line of credit (excluding those requested in the past week)," is the most important feature in the model. The model summary also includes information about the relationship between the input and output of the model. Additionally, from Figure 9-3a, we can observe that higher values of the fraction of all revolving line of credit limits in use increase the non-creditworthy probability.

Another important aspect for researchers is the summarization of the dataset. It involves discovering the intrinsic relationships within the dataset. To this end, we aim to learn a set of simple rules that can summarize the dataset. Tree-based classification approaches can be used to learn such rules. A tree classifier is a set of if-else statements that determines the classification of a data point. As illustrated in Table 9.1, the best optimal tree classifier with our dataset achieves an accuracy of 74.12%. However, the complexity of rules obtained from the best optimal trees classifier makes these rules difficult to analyze and interpret.

To obtain rules that are simple to understand, the decision tree must be constrained to a smaller depth, with fewer features at every node. Consequently, we use a simpler and interpretable optimal tree to obtain more easily analyzed rules. This simple optimal tree is shown in Figure 9-5. Here we show two simple rules that can achieve an accuracy of 72.28%. The rules state that a person will default on a loan (that is, be classified as non-creditworthy) if the number of months since a new line of credit has been requested (excluding those requested in the past week) is greater than 1, and the external risk estimate is less than 75, or if the number of months since a new line of credit has been requested (excluding those requested in the past week) is less than or equal to 1, and the external risk estimate is less than 68.

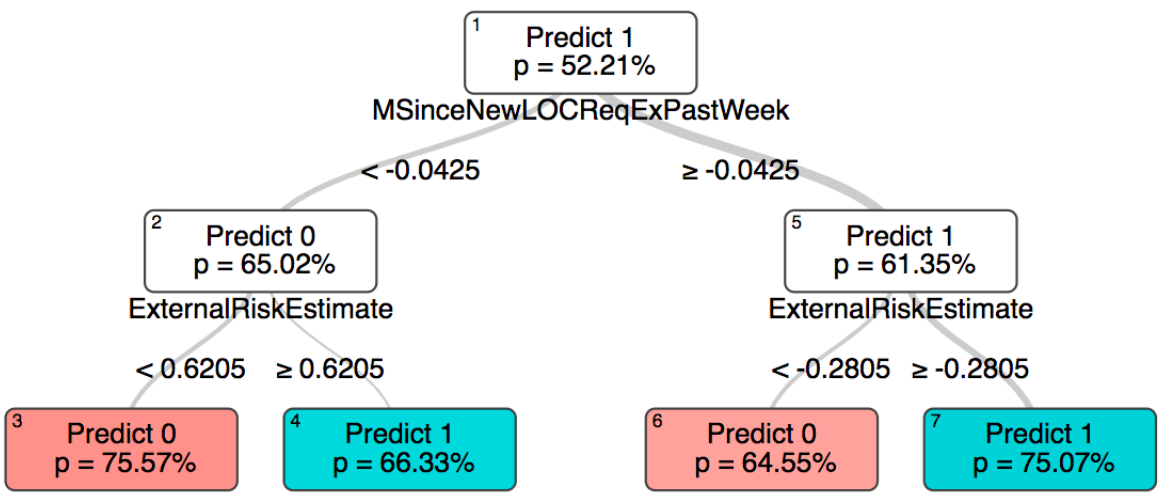


Figure 9-5: Optimal Trees Model
Explainable Optimal Tree Model used to learn simple rules summarizing the dataset. The rules state that a person will default on a loan (that is, be classified as non-creditworthy) if the number of months since a new line of credit has been requested (excluding the past week) is greater than 1 and the external risk estimate is less than 75, or the number of months since a new line of credit has been requested (excluding the past week) is less than or equal to 1 and the external risk estimate is less than 68.

We use inductive logic programming to learn an even simpler set of rules with a small decrease in accuracy compared to the rules generated from the optimal trees model. As discussed earlier, ILP generates rules that are composed of a single condition: for example, using the rule $ExternalRiskEstimate < 72$ to classify people into non-creditworthy and creditworthy groups is able to achieve an accuracy of 70.65%.

In addition to ILP, there exist multiple other rule-finding methods in the literature. However, the ILP approach stands out in terms of its simplicity and ability to learn more effective rules. For example, anchors [240] can be used for generating rules based on data points that are in close proximity. The logistic rule regression/generalized linear rule model described in [284] is another rule-based model, but it leads to multiple rules that are difficult to analyze together. Similarly, the Boolean rule column generation (BRCG) method described in [80] gives a set of rules to describe a dataset, but it requires users to divide feature values into different bins before learning the rules. Data-binning limits the quality of the generated rules, because the binning algorithm may not select the optimal thresholds for the rule.

Using the methods discussed above, data scientists should be more able to summarize datasets and demystify models. Obtaining a good summary of the dataset and an understanding of the model will help in the creation of better classifiers.

9.7 Conclusion

In this paper, we examine several different machine learning models and use suitable tools to create explanations of their respective functions according to the needs of different stakeholders involved in credit risk management. We use state-of-the-art interpretable machine learning techniques, including Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and Diverse Counterfactual Explanations (DiCE), and adapt them to our use case. We demonstrate the importance of domain-specific knowledge in order to explain these black-box models. These domain-specific constraints must be obtained from experts in the field, and they can produce pragmatically valid suggestions and explanations.

Our results demonstrate that, with the right tools, even black-box machine learning models are able to answer a series of important questions for credit risk modeling. These questions include: why does the model classify a data point in a certain way? What small changes in feature values could reverse the model's classification of an individual? How does the model behave in extreme scenarios? What relationships did the model learn? Are the models biased? What is the minimal summary of the dataset?

Answering these questions not only fulfills the legal requirements specified by regulators for the use of machine learning models in credit risk management, but also provides borrowers, lenders, and data scientists with answers to questions that they may desire from ML models. Improvements in the interpretability of ML models can not only accelerate their adoption, but also help domain experts troubleshoot their inner workings, which in turn drives the model to iterate towards higher accuracy and be tailored to its domain.

Many problems in finance and economics have a common mathematical representation and internal statistical structure, and may therefore benefit from our framework to interpret the black-box machine learning models used to analyze them. These include loan defaults, mortgage prepayments, Federal Reserve rate decisions, corporate merger and acquisition decisions, asset return maximization, and insurance claims, among others. Our study is a step in the direction of bridging the gap between these black-box models and their use in a real-world setting.

Future work can extend our explainability analysis by incorporating second-order effects into the explainability algorithms (LIME and DiCE), in order to generate more practical counterfactual suggestions and explanations. It may also be of interest to compare the insights derived from these explainability tools to the traditional factors used for credit risk forecast by involving an expert in the process. Finally, large real-world datasets, in particular those including macroeconomic and demographic features and the size of loan requests, should be used to extend the model's capabilities, in order to help quantify its overall fairness, response to stress testing, and the monetary impact due to the superior performance of black-box machine learning models.

Part V

Evolution of NLP

Chapter 10

From ELIZA to ChatGPT: The Evolution of NLP and Financial Applications

Natural language processing (NLP) has revolutionized the financial industry, providing advanced techniques for the processing, analyzing, and understanding of unstructured financial text. In this Chapter, we provide a comprehensive overview of the historical development of NLP, starting from early rules-based approaches to recent advances in deep-learning-based NLP models. We also discuss applications of NLP in finance along with its challenges, including data scarcity and adversarial examples, and speculate about the future of NLP in the financial industry. To illustrate the capability of current NLP models, we employ a state-of-the-art chatbot as a co-author of this article.

10.1 Introduction

In 1960s, a natural language processing (NLP) program called ELIZA was developed by Professor Joseph Weizenbaum from the Massachusetts Institute of Technology [285], the first chatbot in the history of computer science. It used a simple form of pattern matching to respond to certain keywords and phrases, and although its capabilities were quite limited by today's standards, its impact was unmistakeable. As Weizenbaum explained, "Once my secretary, who had watched me work on the program for many months and therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it, she asked me to leave the room." [64]. In recent years, deep learning techniques have enabled the development of advance chatbots like ChatGPT that generate even more human-like responses. The journey from ELIZA to ChatGPT reflects the ongoing evolution of NLP, as researchers continue to push the boundaries of what is possible with NLP and machine learning.

In the context of finance, the technology company Kensho, a next-generation provider of analytics and artificial intelligence systems to Wall Street banks, was acquired by the financial information powerhouse S&P Global in 2018.[1] The availability of data from S&P Global combined with NLP technology from Kensho saved decades of work-hours of manual data analysis.[2] Other financial companies are currently adopting NLP for the automation of analytics. For example, the investment company BlackRock has designed new NLP models for extracting financial data from unstructured documents.[3] Likewise, many other financial institutions, such as J. P. Morgan, Goldman Sachs, Citigroup, and Vanguard, are adopting NLP for chatbot customer service, equity research, and news analytics among other functions.

The use of NLP in finance has evolved significantly over the past few decades. Initially, NLP was used in finance for comparatively simple tasks, such as sentiment analysis of financial news articles and stock predictions. However, with advances in NLP techniques and increased computational power, the application of NLP to finance has become more sophisticated and diverse. Today, NLP models are used in finance for a wide range of tasks, including asset management, risk management, and impact investing.

The finance industry has unique requirements, and thus unique challenges, for

[1] https://press.spglobal.com/2018-04-09-S-P-Global-Completes-Acquisition-of-Kensho

[2] https://www.waterstechnology.com/data-management/7950146/four-years-on-sps-kensho-buy-yields-new-automation-toools-saving-decades-of-manual-data-analysis

[3] https://www.blackrock.com/aladdin/nlp-aladdin-private-markets-blog

NLP models such as dealing with unstructured financial data and handling the complexities of financial regulations. As a result, the development and implementation of NLP in finance have been slower compared to its use in other industries. However, the potential benefits of using NLP in finance are significant, and include improved risk management, increased operational efficiency and scalability, and enhanced decision making.

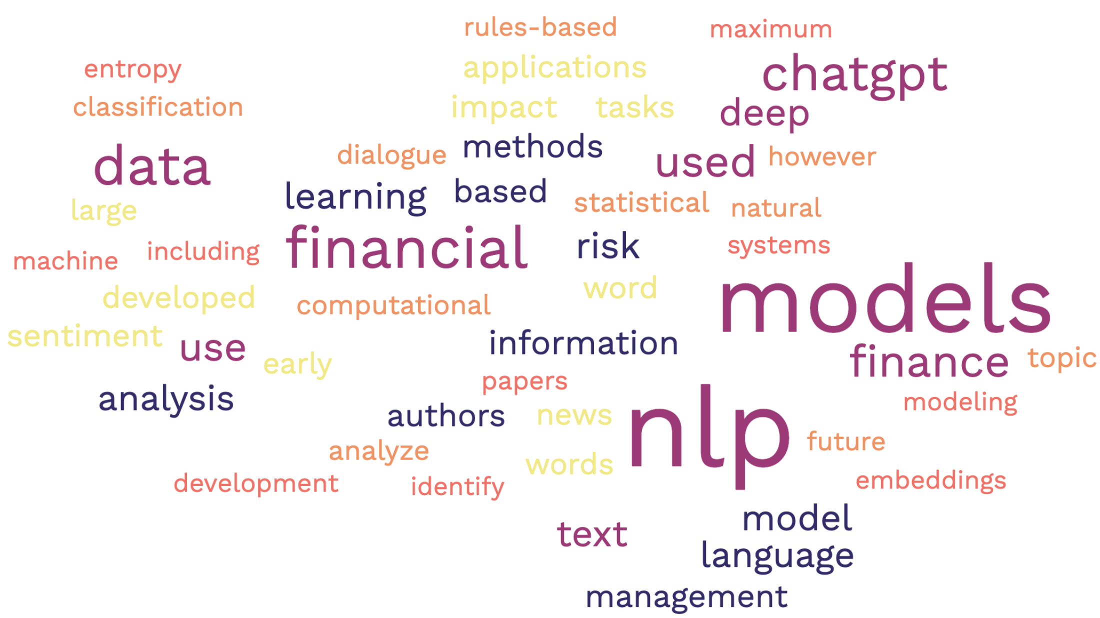


Figure 10-1: Word cloud representing the most frequently occurring words in this article.

While not strictly a survey, this chapter includes a comprehensive overview of the evolution of NLP, starting with the early rules-based models of the 1950s to the 1970s, to the early statistical models of the 1980s and 1990s, followed by the development of embedding-based and advanced statistical models of the 1990s to the 2010s. We then discuss the state-of-the-art deep-learning-based NLP models developed in the 2020s. We explore the applications and current state of NLP in finance, discussing the challenges faced by NLP models and possible insights into their future in this domain. We conclude by proposing opportunities for NLP to transform the financial industry.

To highlight the current state of NLP in finance, we employ a chatbot—OpenAI's ChatGPT—in preparing this article. Although journalistic standards for using chatbots and disclosing such use have not yet been set, we propose to initiate the practice of acknowledging the use of this powerful tool explicitly, and because we include passages generated by ChatGPT verbatim, we have included it as a co-author. The complete interaction between the two human authors and ChatGPT that led to these passages is provided in Appendix.

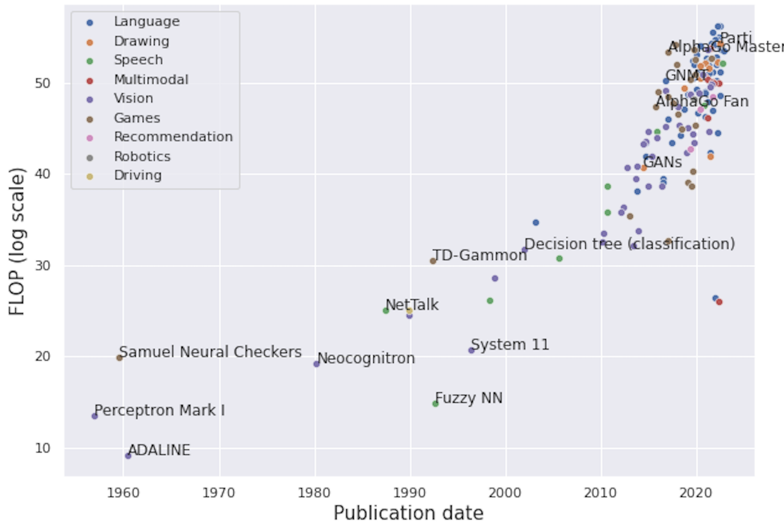
In the next section, we discuss the timeline of development of NLP models, along with a summary of the models. We then discuss the current applications of NLP in finance, following by our view of the future role of NLP in this field.

10.2 Evolution of NLP Models

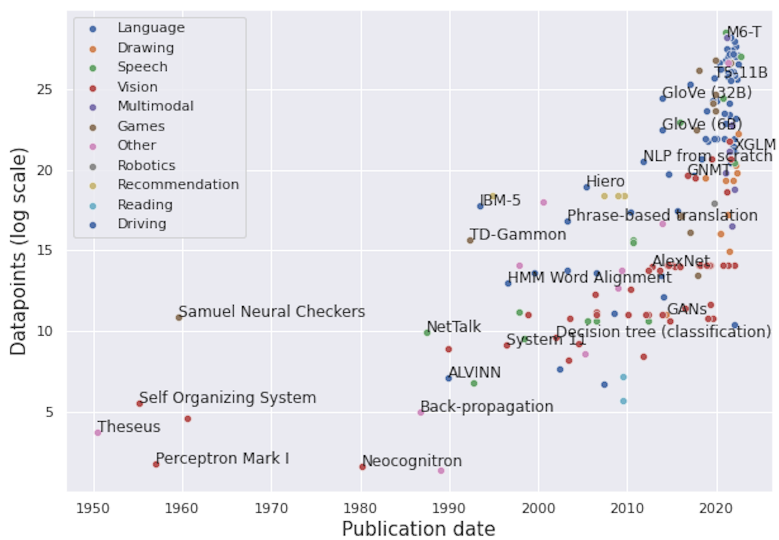
NLP models have evolved over time with theoretical innovation and the availability of data and computational resources. Exhibit 10-2 shows the computation and datapoints used by the artificial intelligence (AI)–including language systems– systems developed over time where we can observe an exponential increase in the computation and datasets used by AI systems. This was possible due to the increase in the computational resources and decreasing cost of computational and data storage facilities.

Increase in computation and datasets implies increasing complexity of the AI/NLP models over time. The earliest NLP-based models in the literature are rules-based systems, developed in the 1950s and 1960s. These systems used handcrafted rules and grammars to analyze natural language text. In the 1980s and 1990s, statistical NLP models emerged, which supplanted rules-based approaches with statistical models. During the 2000s, with the increased availability of datasets and the exponential increase in computational power, machine learning (ML) models were developed that could learn patterns in textual data and make predictions based on this interpretation of text.

In this section, we summarize the models that were developed over this decades-



(**Panel-a**). Computation used to train notable artificial intelligence systems.



(**Panel-b**). Number of datapoints used to train notable artificial intelligence systems.

Figure 10-2: Evolution of computation and number of datapoints for AI systems training.
Dataset is obtained from [253]

long history. We have divided the timeline of NLP models into multiple categories based on model complexity, and consider them one by one.

Rules-Based NLP (1950s to 1970s)

Early NLP research focused on rules-based approaches, including grammar-based parsers and dictionary-based information extraction systems:

- **Grammar-based parsers**: These systems used formal grammars to analyze the structure of text, such as its syntax, and to generate parse trees. This information could then be used to answer questions about the semantics of a text or to perform other NLP tasks, such as information extraction or text classification. Some early influential papers include [63], [242], and [95].

- **Dictionaries**: These systems relied on predefined dictionaries of words and phrases to identify and extract specific pieces of information from text, such as names, dates, and locations [193]. Handcrafted rules were used to classify text into predefined categories, such as positive, negative, or neutral sentiment [231]. They were likewise used to identify and classify named entities in text, such as people, organizations, and locations.

Exhibit 10-3 provides an example of rule-based parsing of a sentence. Such parsing and pattern matching was used by the early chatbots like ELIZA. These rules-based models represented an important early step in the development of NLP, but had several inherent limitations, including the need for extensive handcrafted rules and dictionaries, and the inability to handle exceptions and outliers in text. Some recent works related to rules-based methods are [158], [4], [277] and [220].

Early Statistical NLP Models (1980s to 1990s)

Hidden Markov models (HMMs) [21] and maximum entropy models (MEMs) were among the first statistical NLP models to be developed and applied to NLP tasks,

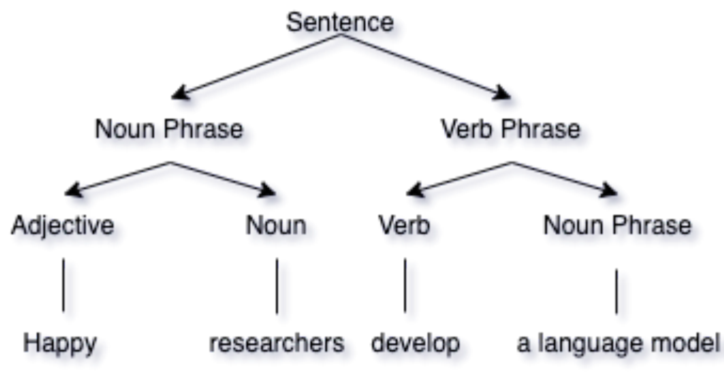


Figure 10-3: Parsing Tree.
Sentence tree constructed by a rule-based grammar parser for the sentence *"Happy researchers develop a language model."*. Such grammar parsers were used by early chatbots like ELIZA to understand the sentence and respond based on the script (dictionary of sentences).

including the tagging of parts of speech, named entity recognition, and speech recognition [18, 29, 28]. HMMs are a type of probabilistic model used for the modeling of sequential data. A sequence is modeled as a series of hidden states and observable outputs. An HMM is then trained using the Baum-Welch algorithm to estimate the parameters that maximize the likelihood of the observed sequence. In an analogous manner, maximum entropy models are probabilistic models that aim to maximize entropy (i.e., uncertainty) subject to certain constraints. MEMs use a log-linear approach to model the probability of an output sequence given an input sequence. However, both these types of models struggled with computational efficiency, and both have known difficulties in modeling complex relationships among variables.

Lexicon-Based and Topic Modeling (1990s to early 2000s)

An important step forward in NLP modeling was the combination of the dictionary and rules-based approach with statistical modeling. These new methods include lexicon-based models and topic modeling, relying on predefined and data-driven lexicons of words and phrases employed with mathematical algorithms to uncover hidden patterns and themes in the data.

Positive	Negative	Uncertain	Litigious
Best	Cyberattack	Vague	Beneficial
Innovativeness	Cybercrime	Anomalous	Bailment
Achieve	Spam	Doubt	Juriousdictionally
Creative	Abolish	Undocumented	Remised
..	..	..	..

(a) Example words from the finance sentiment lexicon developed by [189].
The four sentiments included in the dictionary are positive, negative, uncertain and litigious. Any new document can be given a sentiment score based on the presence of the above words from the dictionary of sentiment words

Topic 1	Topic 2	Topic 3
Loans	Financial Institution	Cashflow
Value	Bank	Interest
Loan	Capital	Liability
Asset	Institution	Income
Tax	Deposit	Rate
..	..	..

(b) Topics learned from 10-K documents using a popular topic modelling algorithm Latent Dirichlet Allocation (LDA).
The topic modelling algorithms doesn't give a topic summary like loans, financial institutions, cashflow as given in the example. These are assigned from the human understanding of the words belonging to a particular topic. Any new sentence (tweet, article, paragraph) can be assigned to a particular topic based on the occurrence of the words in the sentence.
Source: [61]

Table 10.1: Lexicon and topic-modeling examples of NLP .

- **Lexicon-based models**: These models were developed in the early 2000s, and rely on the use of predefined (or constructed from data) dictionaries of words and phrases to analyze and understand natural language text [286]. For example, a lexicon-based model might use a dictionary of financial terms to identify financial concepts in news articles.

- **Topic models**: These models were developed in the late 1990s and early 2000s, and use statistical techniques to identify the main topics discussed in a document or a collection of documents [33, 16]. For example, a topic model might identify market trends or company news as topics from the collection of financial news articles.

The example of financial lexicon model and topic model based on the SEC 10-K filings is presented in Exhibit 10.1. These lexicons and topics are learned from financial documents and are further used to assign sentiment scores or topics in multiple applications. These methods have the significant limitation of relying on the presence of specific words or phrases in a document to determine their content, which does not account for the context in which these words appear. Moreover, these lexicons are predefined or constructed from data, and may not contain all the relevant terms for a given topic, leading to an incomplete analysis.

Embedding-Based NLP (2000s to 2010s)

Word embeddings are a type of NLP technique that represents words as continuous-valued vectors in a high-dimensional mathematical space. These word vectors are able to capture important semantic and syntactic information about words in a way that can be used for various NLP tasks. Some embedding-based methods include the following:

- **Count-based method**: These methods use the frequency of words, phrases, or other linguistic features in a given corpus of text to extract meaningful information, and include techniques such as term frequency-inverse document frequency (TF-IDF), n-grams, and bag-of-words models. Dimensionality reduction, using methods like non-negative matrix factorization and singular value decomposition, can then be applied to count-based embedding to obtain word embeddings.

- **Predictive methods**: These methods, such as GloVe [230] and word2vec [212], use a predictive objective to learn word representations. Their goal is to predict a target word based on surrounding contextual words.

In addition, some methods, such as FastText [150], combine the strengths of both count-based and predictive methods to produce high-quality word embeddings.

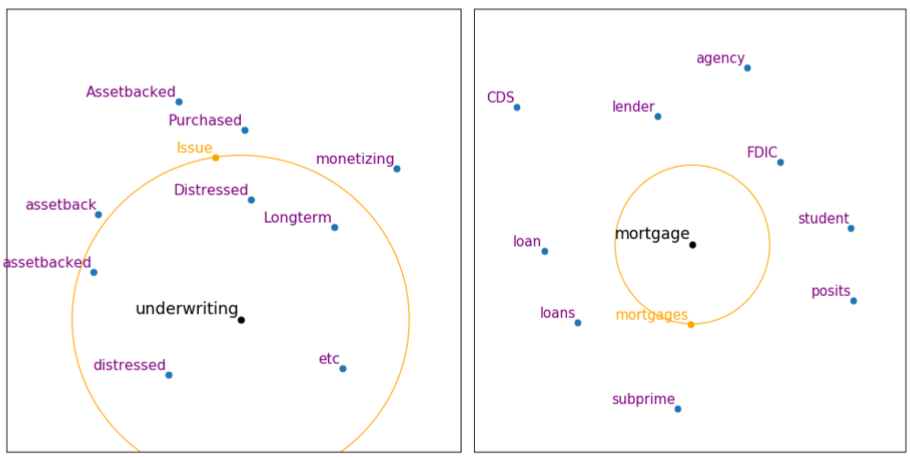


Figure 10-4: The word embedding vectors (obtained using embedding-based NLP method word2vec) plotted on 2-D plane.
The key idea behind the embedding is to encode information and preserve the semantic and relational information. For example, the words in the proximity of underwriting are distressed, assetbacked, longterm which are related to underwriting and likewise for mortgage.
Source: BankFin Embeddings, `https://github.com/sid321axn/bank_fin_embedding`.

As shown in Exhibit 10-4, the the embeddings learn the relational and semantic information. These embeddings can be used for sentiment prediction, classification of sentences etc. However, a disadvantage of embedding-based methods is their dependence on the training corpus and sensitivity to data type and training domain. For example, they often perform poorly on out-of-vocabulary words. Another disadvantage is that word embeddings are context-sensitive. Since the representation for a word can change depending on its context, it can be challenging to use embedding-based methods to effectively capture the meaning of phrases and sentences.

Deep Learning-Based Approaches to NLP (2020s)

Most recently, with the joint exponential growth in data and computational infrastructure, large "deep learning" based language models have been developed and used widely for NLP applications. Initially, recurrent neural networks (RNNs) were developed for the analysis of sequential data, and their use extended to NLP tasks. Along with RNNs, convolution neural networks (CNNs) were used for text classification [157]. The architectures of RNNs and CNNs were developed in 1990s, but they could only be successfully applied with the availability of Big Data and large computational resources. Over time, RNNs were modified for use in applications like neural machine translation, for which sequence to sequence models were developed [268].

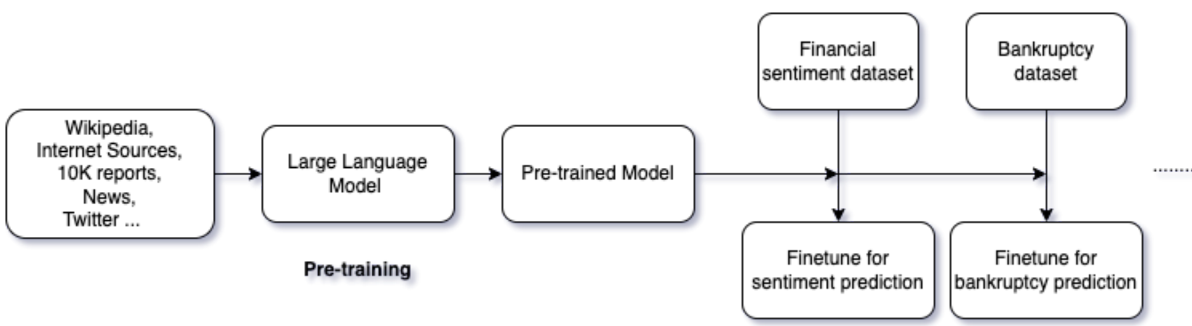


Figure 10-5: Summary of the different steps in the pipeline of a large language model. The first step involves pre-training of the model on the multiple datasets. Once the model is pre-trained, it can be generalized for different tasks like sentiment prediction, banruptcy prediction by finetuning of the model using the dataset specific to the application.

The "imagenet moment" for NLP—referring to the development of generic models that can used for multiple tasks—happened with the development of the transformer-based large language model, specifically the bidirectional encoder representation from transformers (BERT) [85]. The transformer model revolutionized NLP by introducing the self-attention mechanism in NLP, a deep learning technique that allows a model to dynamically weigh the importance of each element in an input sequence. Attention mechanisms can be used in neural network models to understand the relationships between words in a sentence [279]. Over time, more transformer models were developed, including the Generative Pretrained Transformer 2/3 (GPT-2/3) [236], and the robustly optimized BERT approach (RoBERTa) [177]. The detailed survey of large language models can be obtained from [210].

The summary of a large language model training and prediction is included in the Exhibit 10-5. The first step involves pre-training of the model using the variety of big datasets available. It allows the model to learn the general semantics of the language. The pre-trained model can be finetuned for various applications by training them on application specific datasets.[4]

ChatGPT

The most recent development of deep learning-based NLP models is ChatGPT, a chatbot based on the transformer model GPT-3.

- Architecture: ChatGPT is based on the transformer architecture which, as mentioned above, is a deep neural network designed for processing sequential data, such as natural language. The transformer architecture is made up of multiple layers, including an attention mechanism, which allows the model to focus on different parts of the input at different times and dynamically adjust its internal representations.

- Pretraining: Like other GPT models, ChatGPT is pretrained on a large corpus of text data. During pretraining, the model learns to predict the next word in a sentence given its preceding context. This enables the model to build a rich representation of language and to learn patterns and relationships between words and phrases.

- Fine-tuning: Once the model is pretrained, it can be fine-tuned for specific

[4]Finetuning refers to the process of retraining of model by making small adjustments to achieve the desired output.

tasks and domains, such as conversational AI. During fine-tuning, the model is trained on a smaller dataset that is relevant to the target task. This allows the model to learn task-specific information and improve its performance on the target task.

ChatGPT can be used in a variety of conversational AI applications, including in chatbots, virtual assistants, and customer service systems. The model can be used to generate human-like responses to user inputs, answer questions, provide recommendations, and perform other conversational tasks.

A series of chats with ChatGPT is included in Appendix. Based on our experience, we found that ChatGPT is *not* capable of writing a complete manuscript by itself. However, it updates its replies based on the context of the conversation, and is able to understand the fine details about the questions. However, we found that replies from ChatGPT were generally rather long, perhaps due to a bias introduced by long questions and answers in its training data.

Just like any AI or deep learning based model, ChatGPT suffers from limitations. Like all machine learning models, ChatGPT is only as unbiased as the data it is trained upon. If the training data contains biases, such as racial, gender, or cultural biases, the model will reflect those prejudices in its output. Another limitation is ChatGPT's lack of common sense. ChatGPT has been trained on a large corpus of textual data, but it does not have a comprehensive understanding of the world or common-sense knowledge. This can result in unrealistic or nonsensical responses when faced with out-of-domain questions or situations. There are situations of hallucinations as well, when AI models gives incorrect response to the questions with high confidence which may misguide the users. ChatGPT is also not great at logic and reasoning which humans are good at. An example of failed logical reasoning is included in the Appendix Section *Logical Reasoning*. It was able to give a correct response only after a hint. Finally, deploying ChatGPT in real-world applications can be challenging, due to its large size and computational requirements. This can limit its accessibility to smaller organizations or individuals, and make it difficult to deploy in resource-constrained environments.

10.3 Applications

To the best of our knowledge, the earliest use of NLP in finance can be traced to the 1980s, when [114] introduced a methodology for the statistical evaluation of narrative data in accounting reports. Following this application, there were some advances of NLP use in finance in 1990s by [37], which employed data mining techniques for analyzing corporate data, and [290], who performed stock market forecasts from textual web data. However, there was little progress in the use of textual data for financial applications for many years, as early statistical models were not adequate for deep understanding of the domain and the availability of structured data was limited.

From 2010 onward, parallel to the development of social media, the availability of user data increased exponentially, as did the amount of available computational power. This led to the financial community showing a greater interest in data mining from textual data. Since then, the use of NLP in finance has continued to grow and evolve, as companies and researchers have developed more sophisticated NLP models and applied them to a wider range of financial applications.

Today, the use of NLP for financial applications is far from trivial. Existing NLP models have been modified for their specific use in financial applications. For example, [189] developed one of the most popular financial lexicons used to capture sentiment in analysis, and more recently, [153] developed a lexicon for sentiment score generation. Similarly, deep NLP models such as bidirectional encoder representations from transformers (BERT) have been modified and trained on specifically financial data, for example, Financial Phasebank [203] and Reuters TRC2 [173], leading to the development of finBERT [144], which obtained state-of-the-art results for financial sentiment tasks. Today, NLP methods from rules-based to deep NLP models are widely used in finance for a variety of tasks, from sentiment analysis to risk management and algorithmic trading, among others.

A detailed literature review of existing works pertaining to financial NLP is beyond the scope of this article, but in-depth surveys can be found in [292], [112], and [152]. We discuss some of the most important NLP applications in the next few sections.

Risk Management

Risk management in finance refers to the process of identifying, assessing, and controlling risks in financial operations, transactions, and investments. It involves identifying potential sources of financial loss and taking steps to minimize or mitigate their impact on the financial stability and performance of an organization. Key components of financial risk management include market risk, credit risk, operational risk, liquidity risk, and reputational risk. The goal of financial risk management is to minimize the impact of adverse events on an organization's financial performance and stability. It is one of the foremost applications of NLP in finance.

NLP methods are used to extract data from news articles, financial reports, and regulatory filings to identify patterns and relationships between risk factors. [139] performed language analysis to identify event risk, and [43] developed a method to determine risk factors via textual analysis. Exhibit 10-6 shows some example of risk factors obtained using topic modelling. NLP can also be used for regulatory compliance to mitigate financial risk, as it can be used to automatically scan financial documents and identify potential compliance violations, such as insider trading or money laundering. [297] discussed an NLP-based automated compliance checking system. For credit risk, NLP can be used to analyze financial statements and other data to predict a company's credit risk and support the decision-making process in lending and credit management. [135] performed word content analysis for initial public offerings that can assist underwriters in finding more accurate prices.

Similarly, NLP can be used in fraud detection. NLP-based systems can analyze transaction data and detect suspicious patterns that may indicate fraud, helping financial institutions minimize their risk exposure. [190] performed a detailed review on the use of NLP in textual analysis in finance and accounting, while [234] introduced a domain-specific dictionary for predicting fraud from 10-K documents. Similarly, NLP-based network analysis is used to identify interconnections between financial institutions and their exposure to risk. [35] and [143] used NLP in industry classification and identifying interconnections between firms. NLP-based textual analysis is also

Risk Category	Risk Topic Words	Example Companies
Debt Risk	Debt, Payment , Acquisition, Default ..	HEE Equipment services, U S Concrete Inc,
Drug Research Risk	Clinical Trial, Marketing, Approve ..	Hepion Pharmaceuticals, Leap Therapeutics ..
Government Funding Risk	State, Healthcare, Medicare..	Chemed Corp, U S Physical Therapy etc
Advertising Risk	Station, Advertising, Broadcast ..	Beasly broadcast group, Sinclair Broadcast

Figure 10-6: Topic Modelling Risk Factors.
Example of some risk factors that are learned from 10K documents using a topic modelling algorithm along with the words that contribute to the risk topics.
Third column represents the companies that are exposed to the respective risk factors.
Using text based risk factors, we can find the exposure of different risk factors for each company. For example, [188] found that Apple Inc is exposed to Supply chain risk, Software risk, International risk, Internet risk and Demand risk. Such fine-grained risk automated analysis from the textual data is possible due to NLP models.
The source of risk factors is [188].

used for climate risk estimation, which we discuss more fully in the next subsection.

Impact Investing

Impact investing is a form of investment that aims to generate a positive social or environmental impact alongside its financial returns. Impact investments are made by companies, organizations, and funds with the intention to generate a measurable social and environmental impact alongside financial returns. This approach to investment seeks to create financial returns alongside its social or environmental benefits, as opposed to philanthropy, which aims primarily to create social or environmental benefits foremost.

NLP models can be used in impact investing in multiple ways. They can assist in document classification to automatically categorize sustainability reports and disclosures according to specific impact themes, such as climate change, human rights, and biodiversity (e.g., climate annotation in Exhibit 10-7 Panel-a). They can analyze and categorize sustainability-related information contained in a company's annual reports and sustainability disclosures, or perform sentiment analysis of news articles and social media posts to track and quantify the social and environmental impact of companies (e.g., climate risk classification in Exhibit 10-7 Panel-b). Natural language generation can be used to produce impact reports and summaries that highlight a company's positive (or negative) impact on society and the environment. These impact-investing related tasks use NLP models of varying complexity, from dictionary-based methods to find thematic exposure [162] to the deep learning-based program climateBERT [283] for detecting climate context and fact-checking claims of climate risk. Deep learning-based language models can also be used to detect corporate greenwashing (e.g., fact checking in Exhibit 10-7 Panel-c), that is, claims of sustainable activity for marketing purposes without true sustainability efforts [68], a widespread problem in ESG investing.

One concern regarding the use of deep learning-based language models for impact investing is the carbon footprint of these language models, since they are highly computationally intensive. The training of climateBERT caused 115.15 kg of CO2

climate-related classification: Statement	Annotation
ExxonMobil is an American multinational oil and gas company is one of the biggest contributors to global greenhouse gas emissions in the world	Yes
Risk managers need a wide range of risk measures to create a comprehensive picture of risk across their portfolios.	No

(Panel-a). Climate Annotation.
The task is to classify if the statement is climate-related. NLP models can be used for such classification and obtain climate-related documents and statement for the companies. ClimateBERT obtained a F1 score of 98% on the climate annotation task. For dataset information refer [283] and [86]

Climate risk classification: Statement	Annotation
American National Insurance Company recognizes that increased claims activity resulting from catastrophic events and the climate change may also affect the affordability and availability of property and casualty insurance and the pricing for such products.	Risk
A similar approach could be used for allocating emissions in the fossil fuel electricity supply chain between coal miners, transporters and generators.	Neutral

(Panel-b). Climate Sentiment Annotation.
Given a climate-related statement, the aim is to classify if the sentiment of the statement is risk, neutral or opportunistic in climate-related context. ClimateBERT obtained a F1 score of 83% on the above task. For dataset information refer [283] and [86]

Claim	Evidence	Annotation
97% consensus on human-caused global warming has been disproven.	In a 2019 CBS poll, 64% of the US population said that climate change is a “crisis" or a “serious problem", with 44% saying human activity was a significant contributor.	Refute
Global warming is driving polar bears toward extinction	Environmental impacts include the extinction or relocation of many species as their ecosystems change	Support

(Panel-c). Fact Checking.
The aim is to classify, if the evidence supports or refutes the claim. It can be used for detecting greenwashing for the claims made by companies. ClimateBERT obtained a F1 score of 75% on the above task (for dataset information refer [283]). Source of examples: Climate-Fever Dataset[86].

Figure 10-7: Three examples of NLP application to climate-related impact investing.

emissions. An increase in language model research will doubtless have a negative impact on the climate, which goes against the intent of impact investing. However, the information generated through NLP can help impact investors make more informed investment decisions that align with their values, and thus contribute to a greater positive social and environmental impact.

Asset Management

Asset management in finance refers to the professional management of a portfolio of financial assets, such as stocks, bonds, commodities, and real estate. The goal of asset management is to maximize returns and minimize risk for clients, typically by creating and executing investment strategies that are tailored to their specific financial goals and risk tolerance. This typically involves analyzing financial data and market trends, making informed investment decisions, and continuously monitoring and adjusting the portfolio as market conditions change.

NLP has already been used in asset management for various tasks such as sentiment analysis, news analysis, and earnings call analysis. NLP models are used to process and analyze vast amounts of textual data and extract meaningful information to make investment decisions. Sentiment analysis in asset management involves analyzing the tone and emotion of financial news articles and social media posts to predict stock market trends. News analysis involves using NLP to extract relevant information from news articles, such as company news, earnings releases, and analyst reports. This information can be used in the investment decision-making process and improve portfolio management. Earnings call analysis involves using NLP to process and analyze earnings call transcripts. The goal is to extract key financial metrics and sentiment indicators to better understand a company's financial performance when making investment decisions.

Before the adoption of deep learning, a variety of NLP methods were widely used in stock price predictions, including keyword spotting [290], bag-of-words classification [272], manually crafted lexicons [78], pre-existing lexicons [17], mixture models [256], and lexicon-based statistical modeling [153]. They were based on data obtained

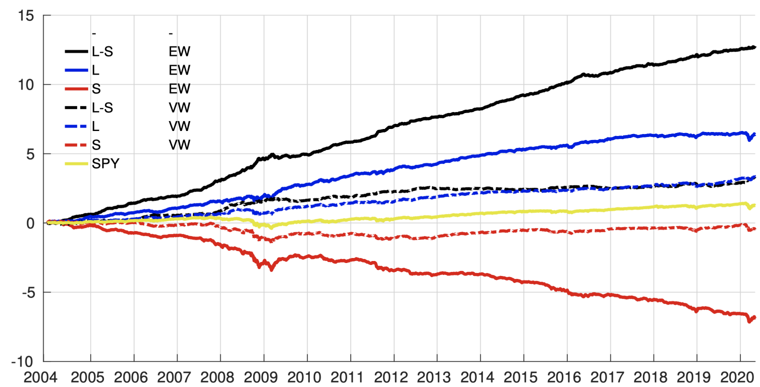


Figure 10-8: Out-of-sample cumulative log returns of portfolio sorted based on sentiment score constructed from news articles.
The sentiment scores are developed based on dictionary based statistical method. L: Long only, S: Short only, L-S: Long-Short, EW: Equally weighted, VW: Value weighted, SPY: S&P500. Long portfolio is constructed by buying 50 stocks with maximum positive sentiment and Short portfolio is constructed by shorting 50 stocks with most negative sentiment. The equal weighted long-short portfolio achieves an annualized sharpe ratio of 4.21 while value weighted long-short portfolio achieves a sharpe of 1.24. The portfolios obtained from the sentiment scores outperforms the baseline of S&P500. Source: [153, Figure 5] ©

from professional periodicals, aggregated news, Twitter, financial reports, and so on. Exhibit 10-8 shows the performance of long-short portfolio (inter-day trading) constructed using sentiment score from [153]. They found that portfolios constructed using dictionary based statistical models can obtain an annualized sharpe ratio of 4.21, implying NLP based sentiment scores can be used for inter-day trading. Similarly, Exhibit 10-9 shows the intra-day event study done by [5] for the Stocktwit and Twitter messages when the sentiment score is above and below three standard deviation of mean. They found that returns are positively correlated to sentiment score and there is increase in liquidity after messages with, high positive and negative sentiment implying NLP based sentiment scores can be used for intra-day trading.

More recently, deep learning-based NLP models have increasingly been supplanting these methods. [152] compared the dictionary-based and deep learning-based approaches (finBERT) for predicting sentiment from earning calls transcripts. NLP methods (i.e., deep learning-based finBERT) applied to Twitter data have been used

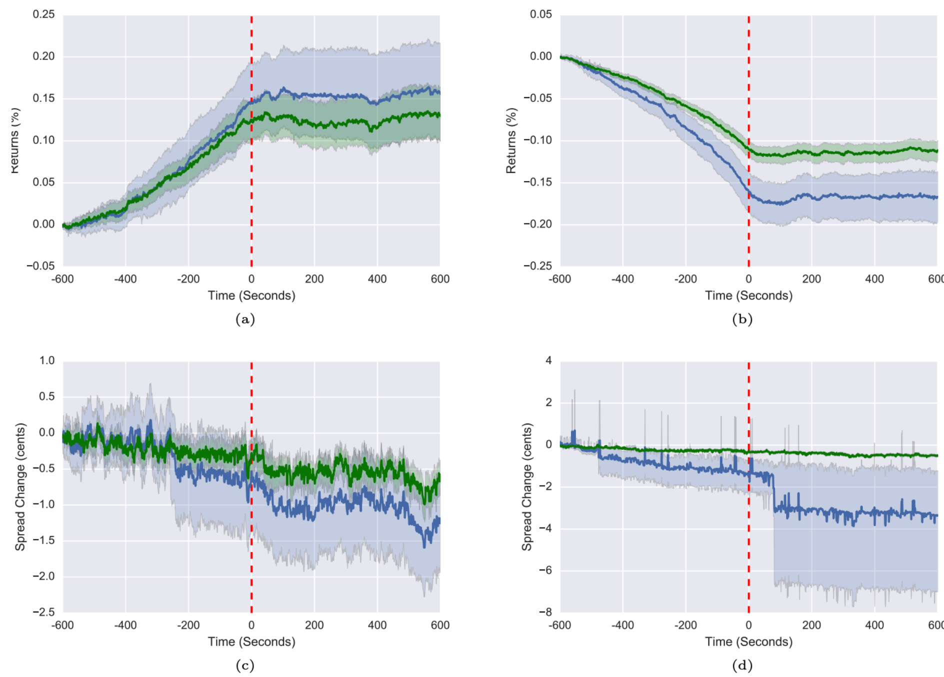


Figure 10-9: Event Study: Stocktwit and Twitter.
Intraday return event studies for Stocktwit and Twitter messages when the sentiment score is above (a) and below (b) three standard deviation of mean, along with spread event study when the sentiment score is above (c) and below (d) three standard deviation of mean. The graphs implies that returns are positively correlated to sentiment score and there is increase in liquidity after messages with large positive and negative sentiment. Source: [5, Exhibit 5].

for the prediction of price movements of cryptocurrencies [299]. The ability of deep models to learn complex patterns which cannot be achieved by linear models makes them highly attractive to the stakeholders involved in asset management.

10.4 What's Next

The future of NLP in finance is promising. Many stakeholders in finance have already been using NLP in their decision-making process, and there are many new and innovative ways that NLP might be used to improve various financial processes in future. However, there are potential limitations. These include:

- **Data Availability, Quality and Bias**: In finance, high quality and relevant data is necessary for NLP models to generate meaningful results. For the applications discussed in earlier sections, good quality Big Data is readily available. However, data is more generally limited and can be unstructured, incomplete or unreliable. Some applications that are affected by data scarcity include sentiment analysis of niche financial products, small business lending, and real estate financing in developing countries. Also, since NLP models are as biased as their data, they can be biased towards or against certain groups or perspectives, leading to discriminatory outcomes and unintended consequences.

- **Complexity of Financial Domains**: The financial domain is complex and constantly evolving, making it challenging for NLP models to keep up with its latest developments and trends. NLP models are often trained on specific datasets in dynamic financial markets, making them vulnerable to overfitting. This limits their ability to generalize to new situations and domains.

- **Regulation and Compliance**: The financial sector is heavily regulated, and NLP models used in finance must comply with various legal and ethical standards, which limit the use and application of NLP technologies. For example, one of the regulatory concern is the black-box nature of deep models. This inherent opacity makes these models hard to interpret to find the reasoning behind the decisions made by these deep NLP models.

- **Adversarial Examples**: Machine learning models are prone to adversarial examples. Adversarial examples in NLP are samples of text that have been deliberately modified to trick an NLP model into making an incorrect prediction. This potential weakness is a vulnerability that could be exploited in real-world settings.

Research is ongoing on the limitations of NLP, and new methods to overcome existing limitations are being developed. For example, explainable AI (xAI) methods to ameliorate the black-box nature of deep learning models like local interpretable model explanations [238] and Shapley values [196] among other methods [77] have been developed and used for financial applications like consumer credit risk [81] and performance attribution [213]. Similarly, these xAI methods can be adopted for use in deep NLP models in finance. Adversarial examples are another widely pursued topic, and researchers have developed ways [11, 294] to train NLP models free of adversarial examples.

A 2017 article in *The Economist* claimed that the "The world's most valuable resource is no longer oil, but data".[5] Big Data is important to financial stakeholders as its availability is necessary to train NLP models. There is significant work on the influence of Big Data in finance [128, 137] and on debiasing methods in deep learning [282, 247] both to make this data unbiased and to make bias-free NLP models more accessible.

The current GPT-3 model has 175 billion parameters. In comparison, the number of neurons in human brain is estimated to be 86 billion. Following popularity of language model, Google Inc. released its conversational AI agent Bard and more advanced NLP models, like GPT-4, are intended to be released in the near future by OpenAI. The future of NLP models thus appears promising. However, NLP models based on historical data will be at a disadvantage in situations when their understand-

[5]https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

ing of the future is different from that of the past (which is often the case in finance). Similarly, the scenarios that require logical reasoning may not be best solved by NLP models. Meanwhile, well-trained humans may have a better understanding of the future and are good at logical reasoning. In such scenarios, humans in the decision-making loop will integrate human expertise and understanding into the data-based machine models. A recent Forbes article calls this integration of human and machine for investing as "quantamental investing," one potential future of investment.[6]

10.5 Conclusion

In this article, we discussed the evolution of NLP models and their applications in finance, from the early rules-based models of the 1950s to the recent deep learning models of the 2020s. We summarized how the models were adopted in finance and discussed their limitations and possibilities for the future.

If "data is the new oil," as many technology mavens have been claiming for some time, then NLP may well be the refinery. The availability of data and NLP tools will significantly alter the financial industry. NLP has already demonstrated its ability to perform tasks such as sentiment analysis, risk management, and impact investing. With advances in NLP technology and the increasing volume of available financial data, NLP is poised to play an increasingly important role in finance, with potential applications ranging from financial analysis and investment decision-making to fraud detection and regulatory compliance.

However, it is important to consider the limitations of NLP, such as the problem of data scarcity, its ethical and legal implications, and the inherent limitations of NLP models themselves. Another important limitation to consider is the apparent "hallucinations" of AI bots which can lead to the strange and sometimes unsettling experiences for users.[7] Overall, the future of NLP in finance is bright, but it will likely be accompanied by a combination of technological pitfalls and practical challengesas

[6]https://www.forbes.com/sites/michaelmolnar/2019/12/12/quantamental-investing-a-fuzzy-term-that-describes-an-inevitable-future/?sh=1a0488c6ff07

[7]https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

the financial industry adapts to these newfound powerful capabilities.

Part VI

Conclusion

A 2017 article in *The Economist* claimed that the "The world's most valuable resource is no longer oil, but data". The data science, machine learning tools can be the refineries that extract meaningful insights from the data. In this thesis chapters, we showed that using right tools and datasets, we can find right peers for industries and crypto-currencies; generate insights and capital for the biopharmaceutical industries investment management; assist in ESG investing; help traders with their stress/activation analytics; assist stakeholders in credit risk management. We hope that the methods introduced will allow the stakeholders to make timely and informed decisions.

There are some challenges with data availability, quality, interpretability of models and bias. However, right tools and right dataset can solve multiple problems.

Appendix A

Peer Grouping : An Artificial Intelligence-Based Industry Peer Grouping System

Data

Dataset Coverage

Table A.1 presents the different datasets and the number of companies in each dataset each year. The starting universe for different datasets is the MSCI USA Investable Market Index (IMI) universe. The index covers approximately 99% of the free-float-adjusted market capitalization in the U.S. For the models trained with all of the datasets, we took the inner join of the companies present in all datasets.

Factor Exposure Features

Table A.2 presents the different factors used in our various models, along with their description. We also include the speed (slow or fast) with which the feature value changes. The slow factors were included at a yearly frequency, and fast factors were included at a monthly frequency.

Year	Returns-Daily	Returns-Monthly	Fundamental Data	10-K Filings	News
2011	2544	2542	2544	2203	2500
2012	2470	2470	2470	2227	2433
2013	2449	2449	2449	2189	2388
2014	2511	2510	2511	2238	2456
2015	2555	2552	2555	2324	2492
2016	2478	2477	2478	2288	2427
2017	2432	2431	2432	2255	2380
2018	2441	2441	2441	2271	2335
2019	2411	2411	2411	2196	2326
2020	2375	2372	–	–	–

Table A.1: Datasets used for feature extraction and the number of companies covered in each dataset every year.

Factor Exposure Description	Speed of Change
One-day reversal	Fast
Historical Beta	Slow
Dividend-to-price ratio	Slow
Analyst-predicted dividend-to-price ratio	Slow
Lower partial moment	Fast
Idiosyncratic lower partial moment	Fast
Hybrid tail covariance risk	Fast
Idiosyncratic hybrid tail covariance risk	Fast
Mean lower partial moment CAPM beta	Fast
Accruals - balance sheet version	Slow
Accruals - cash flow statement version	Slow
Variability in sales	Slow
Variability in earnings	Slow
Variability in cash flows	Slow
Standard deviation of analyst prediction to price	Slow
Enterprise Multiple (EBITDA to EV)	Slow

Trailing Earnings-to-Price Ratio	Slow
Analyst-predicted earnings-to-price ratio	Slow
Long term analyst-predicted growth	Slow
Historical earnings per share growth rate	Slow
Historical sales per share growth rate	Slow
Industry momentum	Fast
Market Leverage	Slow
Book Leverage	Slow
Debt-to-Assets Ratio	Slow
Monthly share turnover	Slow
Quarterly share turnover	Slow
Annual share turnover	Slow
Modified Amihud illiquidity measure	Slow
Pastor-Stambaugh illiquidity measure	Slow
Long-term relative strength	Slow
Long-term historical alpha	Slow
Asset growth	Slow
Issuance growth	Slow
Capital expenditure growth	Slow
Capital expenditure	Slow
Cube of size exposure	Slow
Relative strength	Slow
Asset turnover	Slow
Gross profitability	Slow
Gross margin	Slow
Return on assets	Slow
Return on equity	Slow
Skewness	Slow
Maximum drawdown	Slow
Regional momentum	Fast

Historical Sigma	Slow
Volatility implied by call options, 1 month	Slow
Volatility implied by put options, 1 month	Slow
Volatility implied by call options, 3 month	Slow
Volatility implied by put options, 3 month	Slow
Seasonality	Fast
Revision ratio	Fast
Change in analyst-predicted earnings-to-price	Fast
Change in analyst-predicted earnings per share	Fast
Positive sentiment based on composite sentiment Score	Fast
Positive sentiment based on event sentiment Score	Fast
Sentiment dispersion based on composite sentiment Score	Fast
At-the-money skew	Fast
Short interest	Fast
Short-term reversal	Fast
Log of market capitalization	Slow
Book-to-price ratio	Slow
Sales-to-price ratio	Slow
Cash flow to price ratio	Slow
Structural value	Slow

Table A.2: Description of different factors and their speed of change.

TF-IDF

For featurization of the 10-K filing documents, all documents from 1994 to 2010 were used as training data (corpus). The n-grams of lengths 1 and 2 were considered to learn the vocabulary. Once the vocabulary was learned from all the documents, the documents from 2011 to 2019 were featurized for their use in the models. Each document was represented by an 82606-dimension sparse vector. Given the high dimensionality and sparse nature of TF-IDF features, we performed a truncated SVD to decrease the number of dimensions to 100. The implementation from sklearn and nltk libraries were used.

Distance Metric

In this section, we provide details about the cosine distance model and the ML models not included in the main text.

Cosine Distance

Given feature vectors f_i^k and f_j^k, we determined the cosine distance. i and j indicate the company, and k represents the dataset, such as past returns, 10-K filings, news, or factor exposure. The cosine distance ($d_{i,j}$) between two companies is calculated as $1 - cos(f_i^k, f_j^k)$. When there are multiple datasets, the cosine distance is calculated for all datasets individually and then averaged. We believe that the cosine distance between the companies' features is a strong predictor of the companies' similarity because the cosine distance indicates the relative closeness between a pair of companies.

Machine Learning Models

Ridge Regression: This is a linear ML model widely used for regression tasks. It is a least squares model with an L_2 norm penalty on the weights that increases the generalization power of the model. We used the sklearn implementation of ridge regression. The linearity of the model makes its results interpretable and its input–output relationship transparent.

Neural Network: These are nonlinear supervised learning models used to learn the relationship between input features and the output. They consist of multiple layers that act as a computational engine. In neural networks, the values of the hidden units are computed as

$$h_j(x) = f(w_j + \sum_{i=0}^{n} w_{ij} \cdot x_i)$$

Here, w_{ij} is the weight from input x_i to the hidden unit, h_j. The weights, w_{ij}, can be learned by minimizing the loss function using optimizers like stochastic gradient descent.

The weights of the neural network create complex nonlinear interactions between input features, which makes the decision boundaries learned by neural networks nonlinear. The complex relationships learned in neural networks generally do not have human-interpretable meanings. Hence, the performance improvement obtained by using neural networks comes at the cost of model interpretability. In this work, we used neural networks for regression. We applied the implementation from tensorflow-keras for the construction and training of the model.

XGBoost: XGBoost is a gradient-boosted decision-trees model. A single decision tree is a supervised learning model that learns simple decision rules to segregate data points into different bins. The average of the outputs for all data points reaching the bin is considered as the ultimate output. Each tree is constructed using the classification and regression tree (CART) algorithm. Each tree acts a weak model. XGBoost is based on the idea of combining multiple weak models to create a strong model. Hence, it uses gradient descent to minimize loss while adding new models to the ensemble. Regression trees in general are interpretable models. It is easier to analyze the workings of a single regression tree using if-else conditions; however, XGBoost combines multiple decision trees to make the learned relationship nonlinear. The set of if-else conditions from multiple decision trees is intractable and difficult to interpret.

Dataset Split

Table A.3 gives the details of the training–validation–testing split for the different models trained for various years.

Test Year	Test Inputs	Valid. Year	Valid. Inputs	Train. Year	Train. Inputs
2020	2019	2019	2018	2018	2017
2019	2018	2018	2017	2017	2016
2018	2017	2017	2016	2016	2015
2017	2016	2016	2015	2015	2014
2016	2015	2015	2014	2014	2013
2015	2014	2014	2013	2013	2012
2014	2013	2013	2012	2012	2011

Table A.3: Train-validation-test year split.
The testing year is an out-of-sample year. For example, when we learn the model for grouping companies for the year 2020, the input data (features) is from 2019. For validation, we use the input data from 2018 predicting groupings for 2019, and for training, we use the input data from 2017 predicting groupings for 2018.

Model and Results

Cosine Model Evaluation

Table A.4 reports the average return correlations for the peers obtained from the cosine model. The first row corresponds to the GICS model, and the other rows correspond to cosine models that are based on different datasets. The cosine model performed best when all the datasets were combined, as represented by the row "ALL." However, when compared to GICS, the improvements were minor. Across different datasets, the cosine distance of past returns was the best predictor of future returns correlations, followed by different variants of 10-K features, factors, and news.

After we obtained the cosine similarity, we grouped the companies using hierarchical clustering. Table A.5 presents the average pairwise correlation for clusters obtained using hierarchical clustering on cosine distances. The number of peers (k in top k) is decided based on the number of companies belonging to the cluster at the specified level. The performance of the clusters was comparable to GICS clusters (except for sub-industry groupings). This implies the need for more complex models, such as ML models, to predict the similarity between companies.

Features	Sector	Industry	Sub-industry	Top 10	Top 5	Top 2
GICS	32.40	41.23	43.79	-	-	-
News	27.67	30.15	31.79	33.44	35.23	37.97
Factors	31.45	37.61	39.99	39.80	41.54	44.34
10K-TF-IDF(SVD)	29.87	37.99	42.26	40.97	43.00	45.71
10K-TF-IDF	31.66	40.10	43.09	40.88	42.91	45.76
10K-ALL	31.11	39.49	43.12	41.51	43.49	46.14
Returns-daily	**32.56**	41.12	44.00	43.76	45.86	48.96
ALL	32.55	**41.42**	**44.34**	**44.15**	**46.55**	**49.97**

Table A.4: Evaluation of Cosine Model.
The table presents the average return correlations for peers in the out-of-sample period in which the peers are decided via the cosine model for different datasets. For a fair comparison with GICS, the number of peers for sector, industry, and sub-industry groupings was decided according to the number of companies in the GICS group to which the company belongs. We observed that the performance improvements using all datasets combined (ALL) was small compared to GICS.

Features	Sector	Industry	Sub-industry
GICS	**32.40**	**41.23**	43.79
News	26.55	29.40	31.39
Factors	30.32	36.33	38.70
10K-TF-IDF(SVD)	26.04	36.99	41.55
10K-TF-IDF	30.69	39.18	43.40
10K-ALL	30.62	39.77	42.11
Returns-daily	28.69	38.04	**45.46**
ALL	31.40	39.96	44.05

Table A.5: Results after hierarchical clustering: Cosine Model.
This table presents the average return correlations for the clusters obtained using the cosine model. The number of peers at each level was determined using the number of companies belonging to the cluster at that level of granularity. The cosine model clusters failed to outperform GICS at the sector and industry levels, implying the need of a more complex model, such as ML modeling, to obtain the relative distances between companies.

Performance by Year

In Table A.6, we present the average pairwise return correlation between peer companies for the ML model. For a fair comparison with GICS, the number of peers (k in top k) was decided based on the number of companies belonging to the GICS group of the company at the specified sector, industry, or sub-industry level. The ML models consistently outperformed GICS over all years. The best-performing models are highlighted in bold. The XGBoost model trained on all features, including GICS, was the best-performing model over the years at various levels. We also include a column for the maximum possible average correlation values, calculated by selecting peers on the basis of the forward returns correlation.

In Table A-1, from 2014–2020, we present the average return correlation across all clusters obtained from the different ML models from 2014–2020. The number of peers (k in top k) was decided based on the number of companies belonging to the cluster of the company. The best-performing models over the years are highlighted in bold. The clusters using ML models outperformed the GICS clusters by a large margin at the industry and sub-industry levels.

Bootstrapping

In Table A.6 we observed that the ML models outperformed GICS clusters consistently. We obtained an increase in the forward return correlation up to 4%. To analyze the statistical significance of the improvements, we performed data bootstrapping, and trained a ridge regression (Ridge+GICS) model and calculated the forward returns correlation for the peers predicted by the model. We repeated the process 100 times and obtained the standard errors of the forward returns correlation values for the peers predicted by the model (Table A.7). We observed that errors were much smaller (by an order of magnitude) than the performance improvement obtained. We did not perform bootstrapping for the neural network and XGBoost models owing to the large training time of those models.

Year	Max	GICS	Ridge	Ridge+GICS	NN	NN+GICS	XGB	XGB+GICS
				Sector				
2014	40.33	30.17	34.06	**34.10**	33.57	33.81	34.02	33.78
2015	42.70	32.43	36.45	36.47	36.37	36.82	36.97	**37.08**
2016	43.26	33.91	36.91	36.97	36.97	36.85	37.20	**37.30**
2017	30.92	22.51	23.38	23.25	23.34	23.85	23.75	**23.92**
2018	42.85	33.66	35.79	35.88	36.02	36.09	35.87	**36.42**
2019	40.15	29.79	33.29	33.24	32.97	32.94	33.45	**33.38**
2020	61.87	49.76	53.51	53.54	53.68	53.76	**55.10**	54.09
				Industry				
2014	48.39	38.33	40.77	**40.93**	39.76	40.01	40.40	40.29
2015	52.37	42.79	45.05	45.30	44.91	45.43	45.55	**45.79**
2016	51.97	42.99	44.82	45.00	45.01	44.82	45.33	**45.55**
2017	40.86	32.08	32.87	33.20	32.43	33.29	33.09	**33.45**
2018	51.73	43.06	44.03	44.47	44.37	44.36	44.03	**45.01**
2019	47.71	38.07	39.86	40.01	40.05	40.04	40.30	**40.42**
2020	67.68	56.44	58.82	58.83	58.92	59.13	**59.38**	59.33
				Sub-industry				
2014	51.90	41.53	43.92	**44.10**	42.70	42.91	43.66	43.57
2015	54.39	44.47	46.77	47.06	46.75	47.14	47.10	**47.39**
2016	54.32	45.69	47.30	47.55	47.33	47.12	47.72	**47.90**
2017	44.61	36.08	36.80	**37.25**	36.39	37.19	36.92	37.23
2018	54.14	45.28	46.37	46.81	46.89	46.86	46.70	**47.30**
2019	50.19	41.03	42.40	42.55	42.66	42.70	42.75	**42.97**
2020	67.95	57.93	59.97	60.07	59.94	60.23	60.48	**60.50**

Table A.6: Pairwise Return Correlation: Yearwise.
This table presents the pairwise return correlation over years when the number of peers (top k) is selected by considering the number of peers in the corresponding GICS cluster (for fair comparison with GICS). The best model for each year is bolded.

Year	GICS	Ridge	Ridge+GICS	NN	NN+GICS	XGB	XGB+GICS
				Sector			
2014	30.17	30.49	**30.57**	28.35	29.16	30.12	31.13
2015	32.43	32.27	30.70	29.42	30.78	31.86	**32.69**
2016	33.91	32.15	**34.34**	31.48	31.87	33.56	33.51
2017	**22.51**	19.60	22.04	20.65	20.41	19.24	21.01
2018	33.66	30.79	34.88	30.25	29.05	32.57	**34.66**
2019	29.79	30.11	**30.63**	27.04	24.12	28.10	28.78
2020	49.76	48.89	**51.63**	47.67	49.82	50.22	48.52
				Industry			
2014	38.33	41.12	41.24	31.97	34.54	39.47	**41.44**
2015	42.79	43.70	**45.04**	41.12	41.40	40.77	44.53
2016	42.99	42.95	44.51	41.10	**41.69**	44.36	45.65
2017	32.08	**38.84**	36.26	33.83	34.81	34.80	34.74
2018	43.06	**45.11**	45.10	42.41	44.32	43.98	45.07
2019	38.07	40.85	**41.87**	38.40	40.30	38.78	40.67
2020	56.44	60.14	59.08	58.83	**60.86**	59.30	60.73
				Sub-industry			
2014	41.53	45.36	**45.74**	40.23	39.98	45.57	45.14
2015	44.47	47.87	49.45	46.27	48.74	49.11	**50.93**
2016	45.69	48.60	50.13	49.31	48.18	49.35	**50.21**
2017	36.08	41.47	**42.48**	37.92	39.51	38.62	42.61
2018	45.28	48.61	51.27	50.91	50.63	48.12	**51.86**
2019	41.03	45.35	47.29	47.61	46.73	**48.63**	47.82
2020	57.93	63.95	64.11	65.15	66.00	**66.76**	65.69

Figure A-1: Yearwise Average Pairwise Correlation

This table presents the pairwise return correlation from 2014 to 2020 for GICS clusters and different ML model clusters. The number of peers (k in top k) is decided by the number of companies belonging to the cluster of the company at a specified level. The best model for each year is bolded. For industry and sub-industry clusters, the average pairwise return correlation is higher than for GICS clusters by 10–20%

Year	Sector	Industry	Sub-industry	Top-2	Top-5	Top-10
2014	0.0057	0.0078	0.0097	0.0439	0.0208	0.0157
2015	0.0104	0.0151	0.0160	0.0367	0.0240	0.0206
2016	0.0055	0.0064	0.0077	0.0485	0.0239	0.0163
2017	0.0091	0.0105	0.0110	0.0646	0.00369	0.0286
2018	0.0062	0.0083	0.0080	0.0433	0.0218	0.0159
2019	0.0082	0.0069	0.0061	0.0345	0.0183	0.0149
2020	0.0067	0.0075	0.0077	0.0400	0.0199	0.0144

Table A.7: Return Correlation Standard Errors.
This table presents the standard errors of the correlation in the returns for the peers obtained from the ridge regression model. The standard errors were smaller by orders of magnitude when compared to the improvements obtained. The reported values are in percentages.

Cluster Size

The aim of hierarchical clustering is to avoid having a majority of companies assigned to very few clusters. To evaluate this, we computed the standard deviation of cluster size at the sector, industry, and sub-industry levels, because uniformly distributed clusters will have smaller standard deviations. Table A-2 presents the standard deviation of the cluster-size distribution. It shows that companies are more uniformly distributed in the ML clusters compared to GICS for industry and sub-industry groupings.

Dataset Importance

Figure A-3 presents the importance of different datasets over time. Red indicates higher values,and blue indicates lower ones. From 2014 to 2020, the returns were consistently the most important dataset.

Model:Features	Sector	Industry	Sub-industry
GICS	**124.10**	29.65	16.41
Ridge:ALL	188.95	27.98	15.15
Ridge:ALL+GICS	140.24	26.54	14.89
NN:ALL	159.92	27.34	13.79
NN:ALL+GICS	161.35	26.32	14.47
XGBoost:ALL	124.10	29.65	16.42
XGBoost:ALL+GICS	133.50	**23.37**	**13.50**

Figure A-2: Standard Deviation fo Cluster Sizes.
The table presents the standard deviation of cluster sizes averaged over years. The clusters obtained using ML models were more uniformly distributed than GICS clusters at the industry and sub-industry levels.

	2014	2015	2016	2017	2018	2019	2020
GICS	0.05	0.07	0.08	0.09	0.10	0.14	0.10
News	0.12	0.12	0.12	0.12	0.13	0.15	0.15
10k-D2V	0.18	0.18	0.19	0.19	0.20	0.17	0.19
10k-TFIDF	0.19	0.18	0.22	0.18	0.22	0.20	0.18
Factors	0.24	0.36	0.22	0.25	0.26	0.31	0.37
Returns	0.43	0.39	0.39	0.46	0.45	0.41	0.45

Figure A-3: Feature Importance Datasets.
The above figure presents the importance of features from different datasets aggregated. The returns were consistently the most important dataset from 2014 to 2020.

Appendix B

Peer Grouping: Crypto-Connections:-Exploring the Peers

The Table B.1 presents the id-name and symbol mapping of the cryptocurrencies obtained from CoinGecko.

ID	Name
neo	NEO (NEO)
maker	Maker (MKR)
basic-attention-token	Basic Attention Token (BAT)
iota	IOTA (MIOTA)
tezos	Tezos (XTZ)
eos	EOS (EOS)
tether	Tether (USDT)
cosmos	Cosmos (ATOM)
decentraland	Decentraland (MANA)
bitcoin-cash	Bitcoin Cash (BCH)
binancecoin	Binance Coin (BNB)
filecoin	FileCoin (FIL)
vechain	VeChain (VET)
enjincoin	Enjin Coin (ENJ)

chainlink	Chainlink (LINK)
tron	TRON (TRX)
loopring	Loopring (LRC)
cardano	Cardano (ADA)
kucoin-shares	KuCoin Token (KCS)
theta-token	Theta (THETA)
huobi-token	Huobi Token (HT)
celsius-degree-token	Celsius Network (CEL)
holotoken	Holo (HOT)
nexo	NEXO (NEXO)
ecomi	ECOMI (OMI)
okb	OKX (OKB)
usd-coin	USD Coin (USDC)
bitcoin-cash-sv	Bitcoin SV (BSV)
fantom	Fantom (FTM)
wrapped-bitcoin	Wrapped Bitcoin (WBTC)
quant-network	Quant (QNT)
crypto-com-chain	Cronos (CRO)
theta-fuel	Theta Fuel (TFUEL)
matic-network	Polygon (MATIC)
leo-token	LEO Token (LEO)
harmony	Harmony (ONE)
algorand	Algorand (ALGO)
terra-luna	Terra Classic (LUNC)
ftx-token	FTX Token (FTT)
chiliz	Chiliz (CHZ)
hedera-hashgraph	Hedera Hashgraph (HBAR)
binance-usd	BUSD (BUSD)
blockstack	Stacks (STX)
dai	Dai (DAI)

thorchain	Thorchain (RUNE)
klay-token	Klaytn (KLAY)
kusama	Kusama (KSM)
compound-usd-coin	Compound USD Coin (CUSDC)
compound-ether	Compound Ethereum (CETH)
cdai	Compound Dai (CDAI)
kadena	Kadena (KDA)
solana	Solana (SOL)
arweave	Arweave (AR)
compound-governance-token	Compound Governance Token (COMP)
celo	Celo (CELO)
polkadot	Polkadot (DOT)
avalanche-2	Avalanche (AVAX)
helium	Helium (HNT)
curve-dao-token	Curve DAO Token (CRV)
the-sandbox	The Sandbox (SAND)
sushi	Sushi (SUSHI)
elrond-erd-2	Elrond (EGLD)
terrausd	TerraClassicUSD (USTC)
amp-token	Amp (AMP)
uniswap	Uniswap Protocol Token (UNI)
gala	Gala (GALA)
pancakeswap-token	PancakeSwap (CAKE)
aave	Aave (AAVE)
near	Near (NEAR)
axie-infinity	Axie Infinity Shards (AXS)
ecash	eCash (XEC)
the-graph	The Graph (GRT)
staked-ether	Staked Ether (STETH)
frax	Frax (FRAX)

frax-share	Frax Share (FXS)
flow	Flow - Dapper Labs (FLOW)
olympus	Olympus (OHM)
shiba-inu	Shiba Inu (SHIB)
internet-computer	Internet Computer (ICP)
convex-finance	Convex Finance (CVX)
spell-token	Spell Token (SPELL)
magic-internet-money	Magic Internet Money (MIM)
radix	Radix (XRD)
msol	Marinade Staked SOL (MSOL)
osmosis	Osmosis (OSMO)
huobi-btc	Huobi BTC (HBTC)
bitcoin	Bitcoin (BTC)
ethereum	Ethereum (ETH)
litecoin	Litecoin (LTC)
dash	Dash (DASH)
monero	Monero (XMR)
ethereum-classic	Ethereum Classic (ETC)
dogecoin	Dogecoin (DOGE)
zcash	ZCash (ZEC)
ripple	XRP (XRP)
stellar	Stellar (XLM)
waves	Waves (WAVES)

Table B.1: ID Name Mapping of the Cryptocurrencies.

Appendix C

Finance For Impact: The Reaction of Sponsor Stock Prices to Clinical Trial Outcomes

C.1 Data

The Citeline dataset contains a large number of records about clinical trials performed in the recent past. However, it has significant missing data issues. To use this data for analysis, we perform multiple data cleaning steps, summarized in Figure C-1. Our first step removes all trials that have a missing reason for termination or reasons for termination with an ambiguity in event dates. The termination reasons with ambiguity in their event date include poor enrollment, lack of funding, and business decisions, such as a drug strategy shift or pipeline reprioritization.

From the remaining 46,003 trials, we remove those trials with missing event dates. Since an event study analysis can only be performed for public companies with available stock market returns, we remove the trials that have no stock returns for their sponsors. These sponsor types included academic institutions, government institutions, and private companies.

Next, we exclude those trials from our analysis when confounding factors such as class action lawsuits filed against the company appear in the dataset. Due to the large size of the dataset, it is not possible to isolate all the confounding events. However, some trials are isolated when we are able to find such confounding events. For example, the trials conducted by Repros Therapeutics are not included in our dataset because of the class action lawsuit filed against them on August 8, 2009, that led to abnormal returns around its clinical trial outcome dates.

Finally, we remove those trials with missing clinical trial and sponsor company properties because they are unable to be included in our regression model. After cleaning the dataset, we are left with 13,807 trials for the study. The number of trials across outcomes and trial properties are included in Table C.1.

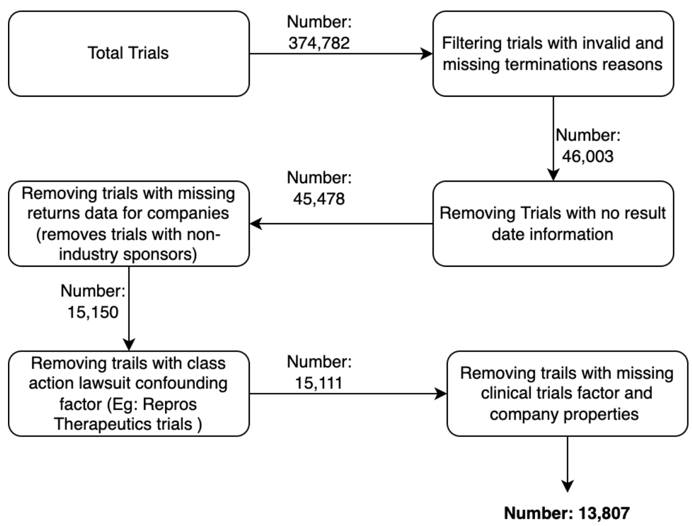


Figure C-1: Data Cleaning Process.
The flowchart presents the different steps involved in data cleaning. The number of trials left in the dataset after every step are given above the arrows. The total number of clinical trials used for the study after data cleaning was 13,807.

C.2 Results: Abnormal Returns Over Time

In this section, we perform the regression given by Eq. (4.1), adding an extra variable for the year on the right-hand side. The year variable is one-hot encoded, taking values from 2001 to 2020. The event dates for the clinical trial were used to obtain the year.

Table C.2 presents the average abnormal returns for different years as determined by the regression. We observed a pattern of increasing abnormal returns during the period of the financial crisis of 2007–2009, and during 2014 to 2017. We found that trials in 2007 had the largest abnormal returns, while trials in 2006 had the lowest abnormal returns.

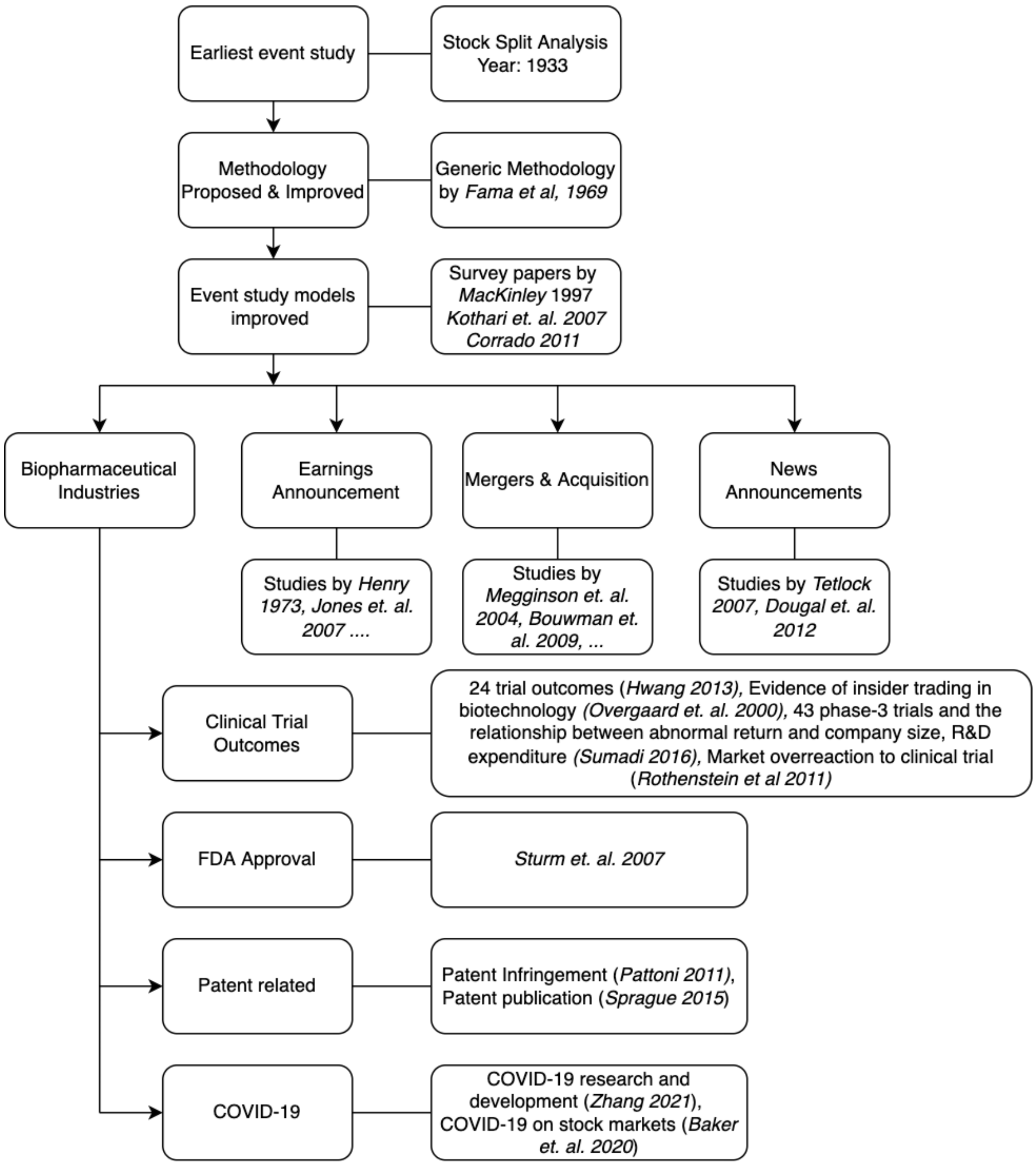


Figure C-2: Literature Review Summary.

Table C.1: Number of Trials

Category	Adverse Effect	Lack of Efficacy	Early +ve Outcome	Primary Endpts. Met	Primary Endpts. not
Total	389	698	127	10156	2437
Phase 1	97	55	12	1060	95
Phase 2	144	364	32	2863	1058
Phase 3	57	168	71	3723	805
Phase 4	18	7	4	1801	297
Phase 1/2	54	77	2	524	99
Phase 2/3	14	26	5	140	63
Phase 3/4	3	0	1	36	18
Big Pharmaceutical	340	568	108	8716	2049
Early-stage Biotechnology	10	28	4	361	105
Late-stage Biotechnology	13	50	11	517	143
Small Pharmaceutical	26	52	4	561	140
Oncology	208	438	73	2626	898
Endocrinology	30	30	11	1815	262
Cardiovascular	33	36	12	869	230
CNS	39	90	14	1431	523
Autoimmune	48	83	10	1815	385
Genitourinary	4	5	0	118	27
Infectious	38	34	6	1454	178
Ophthalmology	3	7	3	177	47
Vaccines	3	6	2	880	58
Randomized	221	420	99	6564	1996
Efficacy	284	612	110	8539	2197
Safety	339	599	114	8903	2021
Placebo Control	132	258	58	3514	1221
Pharmacokinetics	183	308	59	2861	652
Pharmacodynamics	104	189	23	1648	380
Adaptive	10	29	1	96	33
Fixed dose	8	8	0	197	67
Bioequivalence	0	1	0	63	2
Bioavailability	5	0	2	40	2
Cross over	27	25	18	659	164
Drug-drug interaction	6	3	3	53	7
Double-blind/blinded	154	289	73	4760	1433
Open label	239	430	60	5483	1026
Superiority	0	20	15	383	107
Active comparator	53	142	39	2455	667
Non-Inferiority	1	11	1	669	68
Dose Response	46	74	11	878	199
Single Ascending Dose	19	10	2	121	7
Multiple Ascending Dose	12	13	1	145	13
Immunogenicity	23	61	10	1254	157
Observational	4	6	3	564	52
Non-Interventional	1	1	0	227	11
Single Arm	139	234	22	2660	376
Multiple Arm	202	417	88	6167	1800
Basket	1	0	0	10	2
Umbrella	0	0	0	5	5
Patient Registry	1	1	0	34	3

In this table, we present the number of trials across different outcomes and properties. The rows labeled phase 1, 2, 3, 4, 1/2, 2/3, and 3/4 represent the different phases of the clinical trial process, while rows containing biotechnology and pharmaceutical represent the types of the sponsor company. The next nine rows are for the disease therapeutic category of the clinical trial, and the remaining rows are for the clinical trial design. The total number of trials in the dataset is 13,807.

Years	Day 0 (SE)	Day 0-1 (SE)
2001	−0.54 (0.46)	−0.17 (0.64)
2002	0.22 (0.36)	−0.95 (0.51)
2003	−0.73 (0.36)	−0.71 (0.51)
2004	−0.15 (0.23)	−0.38 (0.32)
2005	−0.09 (0.16)	−0.27 (0.23)
2006	−0.56 (0.15)	−1.11 (0.21)
2007	1.64 (0.11)	2.10 (0.16)
2008	−0.10 (0.08)	−0.15 (0.12)
2009	0.21 (0.11)	0.38 (0.16)
2010	−0.90 (0.08)	−0.95 (0.12)
2011	−0.10 (0.07)	0.30 (0.10)
2012	0.00 (0.08)	0.07 (0.11)
2013	−0.12 (0.08)	0.19 (0.11)
2014	0.40 (0.08)	0.61 (0.12)
2015	0.82 (0.08)	1.43 (0.12)
2016	0.99 (0.09)	1.78 (0.13)
2017	0.36 (0.09)	0.54 (0.13)
2018	0.29 (0.10)	−0.06 (0.14)
2019	0.02 (0.11)	−0.13 (0.15)
2020	0.02 (0.13)	−0.17 (0.18)

Table C.2: Regression Coefficient: Time Series.
In this table, we present the regression coefficients for each year included in our dataset. In 2007, 2015, 2016, and 2017, the abnormal returns for clinical trials were higher, while 2006 and 2010 had small abnormal returns for clinical trial outcomes.

Table C.3: Outcome types

Outcome	δ_{-1} (SE)	δ_0 (SE)	δ_{0-1} (SE)	δ_2 (SE)
Early Positive Outcome	0.29 (0.21)	1.99 (0.21)	2.31 (0.30)	-0.01 (0.21)
Lack of Efficacy	0.06 (0.11)	0.68 (0.11)	1.38 (0.15)	-0.23 (0.11)
Primary Endpoints Not Met	0.02 (0.08)	0.87 (0.08)	1.33 (0.11)	0.12 (0.08)
Safety/Adverse Effect	-0.31 (0.11)	0.00 (0.11)	0.06 (0.15)	0.14 (0.11)
Primary Endpoints Met	-0.01 (0.07)	-0.76 (0.07)	-1.08 (0.09)	-0.02 (0.06)
Lack of Efficacy	0.06 (0.09)	1.22 (0.09)	2.03 (0.13)	-0.19 (0.09)
Negative Finding	-0.02 (0.06)	1.29 (0.06)	1.81 (0.09)	0.15 (0.06)
Positive Finding	0.00 (0.05)	-0.16 (0.05)	-0.36 (0.08)	0.01 (0.05)
Closed Early	-0.04 (0.08)	1.35 (0.08)	2.12 (0.11)	-0.07 (0.08)
Full Maturity	0.04 (0.06)	0.59 (0.06)	0.76 (0.09)	0.04 (0.06)

Table C.4: Target Accrual

	θ_{-1} (SE)	θ_0 (SE)	θ_{0-1} (SE)	θ_2 (SE)
Target Accrual	0.023 (0.021)	0.169 (0.021)	0.149 (0.030)	0.026 (0.021)

Table C.5: Company sponsor category

Company Type	γ_{-1} (SE)	γ_0 (SE)	γ_{0-1} (SE)	γ_2
Early-stage Biotech	-0.31 (0.22)	4.61 (0.23)	6.26 (0.32)	0.38 (0.23)
Small Pharma	0.15 (0.17)	1.62 (0.17)	1.79 (0.25)	0.02 (0.17)
Late-stage Biotech	0.30 (0.17)	0.32 (0.17)	0.60 (0.24)	-0.16 (0.17)
Big Pharma	-0.02 (0.10)	-3.66 (0.10)	-4.62 (0.14)	-0.12 (0.10)

Table C.6: Phase categories.

Phase	β_{-1} (SE)	β_0 (SE)	β_{0-1} (SE)	β_2 (SE)
2/3	-0.06 (0.14)	1.47 (0.14)	1.74 (0.20)	0.12 (0.14)
3	0.09 (0.07)	1.08 (0.07)	1.65 (0.09)	-0.03 (0.07)
4	0.01 (0.07)	0.48 (0.07)	0.87 (0.10)	0.08 (0.07)
3/4	-0.14 (0.19)	0.36 (0.19)	0.60 (0.28)	-0.22 (0.19)
2	0.12 (0.07)	0.31 (0.07)	0.48 (0.09)	0.12 (0.07)
1	0.2 (0.09)	-0.38 (0.09)	-0.32 (0.13)	0.10 (0.09)
1/2	-0.04 (0.11)	-0.64 (0.11)	-1.03 (0.16)	0.01 (0.11)

Table C.7: Disease therapeutic category

Disease	η_{-1} (SE)	η_0 (SE)	η_{0-1} (SE)	η_2 (SE)
Genitourinary	0.07 (0.27)	1.60 (0.28)	3.44 (0.39)	0.17 (0.28)
Ophthalmology	0.12 (0.19)	0.32 (0.19)	0.75 (0.27)	0.25 (0.19)
Vaccines (Infectious Disease)	-0.04 (0.13)	-0.07 (0.13)	0.33 (0.19)	0.17 (0.13)
CNS	0.10 (0.13)	0.23 (0.13)	-0.12 (0.18)	0.15 (0.13)
Oncology	-0.06 (0.12)	-0.18 (0.12)	-0.40 (0.17)	0.01 (0.12)
Cardiovascular	0.13 (0.12)	0.07 (0.12)	-0.52 (0.17)	0.04 (0.12)
Infectious Disease	0.07 (0.11)	-0.47 (0.11)	-0.61 (0.15)	-0.09 (0.11)
Autoimmune/Inflammation	0.04 (0.12)	-0.31 (0.12)	-0.71 (0.17)	0.01 (0.12)
Metabolic/Endocrinology	-0.03 (0.11)	-0.52 (0.11)	-0.96 (0.16)	0.08 (0.11)

Table C.8: Clinical Trial Design

Design	ζ_{-1} (SE)	ζ_0 (SE)	ζ_{0-1} (SE)	ζ_2 (SE)
Placebo Control	-0.03 (0.08)	1.12 (0.08)	1.38 (0.11)	0.11 (0.08)
Pharmacokinetics	-0.02 (0.07)	0.50 (0.07)	1.0 (0.1)	-0.07 (0.07)
Adaptive	0.53 (0.26)	0.48 (0.26)	0.80 (0.37)	-0.34 (0.26)
Safety	-0.09 (0.07)	0.60 (0.07)	0.71 (0.09)	0.25 (0.07)
Fixed Dose	-0.03 (0.16)	0.52 (0.16)	0.68 (0.23)	0.12 (0.16)
Active Comparator	0.14 (0.07)	0.09 (0.07)	0.41 (0.09)	0.11 (0.07)
Non Iinterventional	-0.05 (0.14)	0.48 (0.13)	0.32 (0.19)	-0.27 (0.13)
Multiple Arm	0.02 (0.09)	0.20 (0.09)	0.17 (0.13)	-0.08 (0.09)
Observational	0.02 (0.11)	0.11 (0.11)	0.07 (0.16)	0.09 (0.11)
Randomized	-0.11 (0.09)	0.14 (0.09)	0.02 (0.14)	-0.05 (0.09)
Single Arm	-0.08 (0.10)	0.14 (0.10)	-0.02 (0.14)	-0.11 (0.1)
Efficacy	-0.06 (0.07)	-0.19 (0.07)	-0.08 (0.10)	-0.05 (0.07)
Immunogenicity	-0.03 (0.09)	0.24 (0.09)	-0.16 (0.12)	0.04 (0.09)
Dose Response	-0.01 (0.1)	-0.05 (0.1)	-0.21 (0.14)	-0.05 (0.1)
Cross Over	-0.11 (0.11)	-0.08 (0.11)	-0.27 (0.15)	-0.10 (0.11)
Superiority	-0.02 (0.11)	-0.14 (0.11)	-0.28 (0.16)	0.08 (0.11)
Open Label	0.07 (0.09)	-0.06 (0.09)	-0.42 (0.13)	-0.05 (0.09)
Double blind/blinded	0.03 (0.11)	-0.10 (0.11)	-0.46 (0.15)	-0.19 (0.11)
Pharmacodynamics	0.01 (0.07)	-0.60 (0.08)	-0.77 (0.11)	0.10 (0.07)
Non Inferiority	-0.35 (0.1)	-0.68 (0.10)	-0.93 (0.14)	-0.01 (0.1)
Single Ascending Dose	0.13 (0.28)	-0.41 (0.28)	-0.96 (0.39)	0.28 (0.28)
Multiple Ascending Dose	-0.12 (0.25)	-1.86 (0.25)	-1.90 (0.36)	-0.13 (0.25)

Table C.9: Regression Constant

	α_{-1} (SE)	α_0 (SE)	α_{0-1} (SE)	α_2 (SE)
Constant	0.04 (0.13)	2.78 (0.13)	4.01 (0.18)	-0.01 (0.13)

In Tables C.3 to C.9, we present the regression coefficients obtained using Eq. (4.1) for the abnormal returns (computed using **constant mean** model) for the day prior to the event date (-1), on the day of the event (0 and 0–1), and the day after the event date (2). The rows are the trial properties and columns are the coefficients for the days around the event(-1, 0, 0-1 , 2). The coefficients can be interpreted as the average abnormal returns observed for the sponsor company for a trial outcome with a specific property, controlling for other properties of the trial. The abnormal returns are statistically insignificant prior to and after the event date (day -1 and day 2), while they are significant for properties on the days of the event (day 0 and day 0-1).

Table C.10: Outcome types

Outcome	δ_{-1} (SE)	δ_0 (SE)	δ_{0-1} (SE)	δ_2 (SE)
Early Positive Outcome	0.14 (0.17)	1.59 (0.17)	1.83 (0.24)	-0.02 (0.17)
Primary Endpoints Not Met	-0.03 (0.06)	0.9 (0.07)	1.35 (0.09)	0.11 (0.06)
Lack of Efficacy	0.08 (0.09)	0.62 (0.09)	1.34 (0.13)	-0.22 (0.09)
Safety/Adverse Effect	-0.17 (0.09)	0.21 (0.09)	0.22 (0.13)	0.12 (0.09)
Primary Endpoints Met	-0.01 (0.05)	-0.57 (0.06)	-0.81 (0.08)	0.03 (0.05)
Lack of Efficacy	0.08 (0.08)	1.11 (0.08)	1.91 (0.11)	-0.19 (0.08)
Negative Finding	-0.05 (0.05)	1.29 (0.05)	1.76 (0.08)	0.14 (0.05)
Positive Finding	-0.01 (0.05)	-0.04 (0.05)	-0.19 (0.07)	0.06 (0.05)
Closed Early	0.0 (0.07)	1.32 (0.07)	2.05 (0.09)	-0.07 (0.07)
Full Maturity	0.0 (0.06)	0.66 (0.06)	0.86 (0.08)	0.08 (0.06)

Table C.11: Target Accrual

	θ_{-1} (SE)	θ_0 (SE)	θ_{0-1} (SE)	θ_2 (SE)
Target Accrual	0.017 (0.016)	0.171 (0.016)	0.154 (0.023)	0.015 (0.016)

Table C.12: Company sponsor category

Company Type	γ_{-1} (SE)	γ_0 (SE)	γ_{0-1} (SE)	γ_2
Early-stage Biotech	-0.37 (0.22)	4.52 (0.22)	6.42 (0.31)	0.29 (0.22)
Small Pharma	0.13 (0.16)	1.71 (0.16)	1.77 (0.23)	0.13 (0.16)
Late-stage Biotech	0.32 (0.16)	0.33 (0.16)	0.54 (0.22)	-0.17 (0.16)
Big Pharma	-0.01 (0.09)	-3.65 (0.09)	-4.68 (0.13)	-0.11 (0.09)

Table C.13: Phase categories.

Phase	β_{-1} (SE)	β_0 (SE)	β_{0-1} (SE)	β_2 (SE)
3/4	-0.10 (0.12)	1.45 (0.12)	1.83 (0.17)	0.13 (0.12)
3	0.05 (0.06)	1.03 (0.06)	1.60 (0.08)	-0.02 (0.06)
3/4	-0.15 (0.16)	0.35 (0.16)	0.74 (0.23)	-0.08 (0.16)
4	-0.04 (0.06)	0.39 (0.06)	0.69 (0.08)	0.06 (0.06)
2	0.04 (0.06)	0.27 (0.06)	0.39 (0.08)	0.15 (0.06)
1	0.16 (0.08)	-0.33 (0.08)	-0.38 (0.12)	0.04 (0.08)
1/2	-0.11 (0.1)	-0.55 (0.10)	-1.00 (0.15)	-0.04 (0.11)

Table C.14: Disease therapeutic category

Disease	η_{-1} (SE)	η_0 (SE)	η_{0-1} (SE)	η_2 (SE)
Genitourinary	0.11 (0.23)	1.63 (0.23)	3.62 (0.32)	0.20 (0.23)
Ophthalmology	0.16 (0.16)	-0.10 (0.16)	0.40 (0.23)	0.21 (0.16)
Vaccines (Infectious Disease)	0.02 (0.12)	-0.19 (0.12)	0.35 (0.17)	0.11 (0.12)
CNS	0.1 (0.11)	0.27 (0.11)	0.16 (0.16)	0.06 (0.11)
Oncology	0.03 (0.11)	-0.22 (0.11)	-0.26 (0.15)	-0.06 (0.11)
Cardiovascular	0.14 (0.11)	-0.05 (0.11)	-0.53 (0.15)	0.02 (0.11)
Autoimmune/Inflammation	0.12 (0.11)	-0.42 (0.11)	-0.59 (0.16)	-0.04 (0.11)
Infectious Disease	0.07 (0.09)	-0.54 (0.09)	-0.66 (0.13)	-0.13 (0.09)
Metabolic/endocrinology	0.06 (0.1)	-0.58 (0.1)	-0.87 (0.14)	0.09 (0.1)

Table C.15: Clinical Trial Design

Design	ζ_{-1} (SE)	ζ_0 (SE)	ζ_{0-1} (SE)	ζ_2 (SE)
Placebo Control	-0.09 (0.07)	1.11 (0.07)	1.23 (0.1)	0.15 (0.07)
Adaptive	0.46 (0.24)	0.63 (0.24)	1.0 (0.34)	-0.43 (0.24)
Pharmacokinetics	-0.04 (0.06)	0.45 (0.06)	1.0 (0.09)	-0.09 (0.06)
Fixed Dose	-0.01 (0.14)	0.6 (0.14)	0.7 (0.2)	0.04 (0.14)
Safety	-0.05 (0.06)	0.6 (0.06)	0.69 (0.08)	0.21 (0.06)
Active Comparator	0.07 (0.06)	0.14 (0.06)	0.48 (0.08)	0.09 (0.06)
Non Iinterventional	0.01 (0.11)	0.44 (0.11)	0.38 (0.16)	-0.29 (0.11)
Multiple Arm	0.03 (0.08)	0.17 (0.08)	0.18 (0.11)	-0.04 (0.08)
Observational	0.01 (0.09)	0.13 (0.09)	0.05 (0.13)	0.0 (0.09)
Randomized	-0.1 (0.08)	0.13 (0.08)	-0.06 (0.12)	-0.09 (0.08)
Efficacy	-0.03 (0.06)	-0.19 (0.06)	-0.1 (0.09)	-0.08 (0.06)
Single Arm	-0.06 (0.09)	0.06 (0.09)	-0.1 (0.13)	-0.03 (0.09)
Immunogenicity	-0.01 (0.08)	0.29 (0.08)	-0.16 (0.11)	0.06 (0.08)
Cross Over	-0.06 (0.09)	-0.09 (0.09)	-0.3 (0.13)	-0.06 (0.09)
Superiority	-0.02 (0.1)	-0.21 (0.1)	-0.36 (0.14)	0.09 (0.1)
Double blind/blinded	0.06 (0.09)	-0.07 (0.09)	-0.36 (0.13)	-0.17 (0.09)
Dose Response	0.05 (0.09)	-0.13 (0.09)	-0.36 (0.12)	-0.02 (0.09)
Open Label	0.05 (0.08)	0.0 (0.08)	-0.4 (0.11)	-0.03 (0.08)
Pharmacodynamics	0.0 (0.07)	-0.61 (0.07)	-0.77 (0.09)	0.09 (0.07)
Non Inferiority	-0.35 (0.09)	-0.7 (0.09)	-0.97 (0.12)	-0.04 (0.08)
Single Ascending Dose	0.13 (0.26)	-0.57 (0.26)	-1.13 (0.37)	0.41 (0.26)
Multiple Ascending Dose	-0.16 (0.23)	-1.71 (0.23)	-1.66 (0.33)	-0.14 (0.23)

Table C.16: Regression Constant

	α_{-1} (SE)	α_0 (SE)	α_{0-1} (SE)	α_2 (SE)
Constant	0.02 (0.12)	2.75 (0.12)	3.93 (0.17)	0.03 (0.12)

In Tables C.10 to C.16, we present the regression coefficients obtained using Eq. (4.1) for the abnormal returns (computed using **market** model) for the day prior to the event date (-1), on the day of the event (0 and 0–1), and the day after the event date (2). Rows are the properties in each table, and columns are the coefficients for the days around the event(-1, 0, 0-1 , 2). The coefficients can be interpreted as the average abnormal returns observed for the sponsor company for a trial outcome with a specific property, controlling for other properties of the trial. The abnormal returns are statistically insignificant prior to and after the event date (day -1 and day 2) while they are significant for properties on the days of the event (day 0 and day 0-1).

Table C.17: Outcome types

Outcome	δ_{-1} (SE)	δ_0 (SE)	δ_{0-1} (SE)	δ_2 (SE)
Primary Endpoints not Met (Negative Outcome)	0.01 (0.06)	1.78 (0.06)	2.64 (0.08)	0.08 (0.06)
Primary Endpoints Met (Positive Outcome)	0.02 (0.05)	0.24 (0.05)	0.31 (0.07)	-0.02 (0.05)

Table C.18: Target Accrual

	θ_{-1} (SE)	θ_0 (SE)	θ_{0-1} (SE)	θ_2 (SE)
Target Accrual	0.020 (0.016)	0.226 (0.016)	0.217 (0.023)	0.009 (0.017)

Table C.19: Company sponsor category

Company Type	γ_{-1} (SE)	γ_0 (SE)	γ_{0-1} (SE)	γ_2
Early-stage Biotech	-0.29 (0.21)	4.24 (0.21)	6.03 (0.29)	0.33 (0.21)
Small Pharma	0.12 (0.17)	1.75 (0.16)	1.40 (0.23)	0.24 (0.16)
Late-stage Biotech	0.27 (0.16)	-0.43 (0.16)	-0.17 (0.22)	-0.21 (0.16)
Big Pharma	0.00 (0.09)	-3.46 (0.09)	-4.33 (0.13)	-0.15 (0.09)

Table C.20: Phase categories.

Phase	β_{-1} (SE)	β_0 (SE)	β_{0-1} (SE)	β_2 (SE)
2/3	-0.12 (0.13)	1.80 (0.13)	1.99 (0.18)	0.00 (0.13)
3	0.01 (0.06)	0.67 (0.06)	1.12 (0.08)	0.00 (0.06)
2	0.06 (0.06)	0.23 (0.06)	0.41 (0.08)	0.11 (0.06)
4	-0.02 (0.06)	0.10 (0.06)	0.29 (0.08)	0.04 (0.06)
3/4	-0.09 (0.16)	0.02 (0.16)	0.22 (0.22)	-0.03 (0.15)
1	0.16 (0.09)	-0.32 (0.09)	-0.25 (0.13)	0.05 (0.09)
1/2	-0.1 (0.12)	-0.79 (0.12)	-1.17 (0.16)	0.07 (0.12)

Table C.21: Disease therapeutic category

Disease	η_{-1} (SE)	η_0 (SE)	η_{0-1} (SE)	η_2 (SE)
Genitourinary	0.05 (0.23)	1.67 (0.23)	3.49 (0.32)	0.36 (0.23)
Ophthalmology	0.08 (0.16)	0.21 (0.16)	0.79 (0.23)	0.19 (0.16)
Vaccines (Infectious Disease)	0.06 (0.12)	-0.13 (0.12)	0.19 (0.17)	0.18 (0.12)
CNS	0.00 (0.12)	0.24 (0.12)	-0.14 (0.17)	0.16 (0.12)
Infectious Disease	-0.07 (0.09)	-0.67 (0.10)	-0.85 (0.14)	-0.04 (0.10)
Oncology	-0.05 (0.11)	-0.52 (0.11)	-0.89 (0.16)	0.08 (0.11)
Autoimmune/Inflammation	0.04 (0.11)	-0.56 (0.11)	-0.94 (0.16)	0.06 (0.11)
Cardiovascular	0.00 (0.11)	-0.33 (0.11)	-0.94 (0.15)	0.14 (0.11)
Metabolic/Endocrinology	0.01 (0.10)	-0.72 (0.10)	-1.18 (0.14)	0.17 (0.10)

Table C.22: Clinical Trial Design

Design	ζ_{-1} (SE)	ζ_0 (SE)	ζ_{0-1} (SE)	ζ_2 (SE)
Placebo Control	-0.08 (0.07)	1.11 (0.07)	1.37 (0.10)	0.04 (0.08)
Pharmacokinetics	-0.04 (0.06)	0.59 (0.06)	0.92 (0.09)	-0.03 (0.06)
Safety	-0.05 (0.06)	0.48 (0.06)	0.70 (0.08)	0.04 (0.06)
Fixed Dose	0.00 (0.13)	0.39 (0.13)	0.57 (0.18)	0.12 (0.13)
Active Comparator	0.06 (0.06)	0.19 (0.06)	0.37 (0.08)	0.14 (0.06)
Non Iinterventional	0.02 (0.11)	0.39 (0.11)	0.33 (0.15)	-0.30 (0.11)
Observational	0.02 (0.09)	0.25 (0.09)	0.18 (0.13)	0.04 (0.09)
Multiple Arm	0.06 (0.08)	0.17 (0.08)	0.16 (0.12)	0.04 (0.08)
Open Label	0.07 (0.08)	0.22 (0.08)	-0.07 (0.12)	-0.04 (0.08)
Efficacy	-0.05 (0.06)	-0.12 (0.06)	-0.07 (0.09)	-0.05 (0.06)
Randomized	-0.09 (0.09)	0.06 (0.09)	-0.11 (0.12)	-0.14 (0.09)
Dose Response	0.01 (0.09)	0.02 (0.09)	-0.13 (0.13)	-0.06 (0.09)
Single Arm	-0.04 (0.09)	0.06 (0.09)	-0.13 (0.13)	0.00 (0.09)
Superiority	-0.03 (0.10)	-0.03 (0.10)	-0.14 (0.14)	0.02 (0.10)
Adaptive	0.41 (0.27)	0.06 (0.26)	-0.16 (0.38)	-0.54 (0.27)
Double blind/blinded	0.08 (0.10)	0.13 (0.09)	-0.26 (0.13)	-0.05 (0.10)
Immunogenicity	-0.01 (0.08)	0.10 (0.08)	-0.35 (0.11)	0.00 (0.08)
Cross Over	0.01 (0.10)	-0.40 (0.10)	-0.57 (0.14)	-0.03 (0.10)
Single Ascending Dose	0.01 (0.29)	-0.07 (0.28)	-0.60 (0.40)	0.52 (0.29)
Non Inferiority	-0.37 (0.08)	-0.60 (0.08)	-0.73 (0.12)	-0.05 (0.08)
Pharmacodynamics	0.0 (0.07)	-0.73 (0.07)	-0.83 (0.1)	0.00 (0.07)
Multiple Ascending Dose	-0.08 (0.24)	-1.99 (0.24)	-2.31 (0.34)	-0.13 (0.24)

Table C.23: Regression Constant

	α_{-1} (SE)	α_0 (SE)	α_{0-1} (SE)	α_2 (SE)
Constant	0.03 (0.10)	2.01 (0.10)	2.95 (0.14)	0.06 (0.10)

In Tables C.17 to C.23, we present the regression coefficients obtained using Eq. (4.1) for the abnormal returns (computed using the Fama-French 5-factor model) for the day prior to the event date (-1), on the day of the event (0 and 0–1), and the day after the event date (2). The abnormal return attribution model is built using a subset of dataset, including only two outcomes: primary endpoints met (positive outcome) and primary endpoints not met (negative outcome). Rows are the properties in each table and columns are the coefficients for the days around the event (-1, 0, $0 - 1$, 2). The coefficients can be interpreted as the average abnormal returns observed for the sponsor company for a trial outcome with that specific property, controlling for other properties of the trial. The abnormal returns are statistically insignificant prior to and after the event date (day -1 and day 2), while they are significant for properties on the days of the event (day 0 and day 0-1).

Appendix D

Appendix Megafund

D.1 Parameters of Simulation

Construction of a drug development portfolio involved multiple parameters. We describe them in this section along with the values used for the simulation.

Probability of Success

One of the parameter is the probability of a successful phase transition of the trials. We estimated them from the [228], [138], [274] and [288]. The PoS used in the simulation are included in the Figure D.1.

	Probability of Success (%)					
Therapeutic Area	Preclinical to Phase 1	Phase 1 to Phase 2	Phase 2 to Phase 3	Phase 3 to NDA	NDA to Approval	Preclinical to Approval
Oncology	69.0	64.1	30.1	40.0	82.1	4.4
Metabolic/Endocrinology	69.0	63.8	45.6	63.1	84.4	10.7
Cardiovascular	69.0	64.2	40.2	54.9	84.4	8.2
CNS	69.0	63.3	37.7	53.4	83.9	7.4
Autoimmune/Inflammation	69.0	68.0	39.5	62.5	83.2	9.6
Genitourinary	69.0	66.3	46.3	68.4	83.1	12.0
Infectious Disease	69.0	67.3	51.5	64.6	86.8	13.4
Ophthalmology	69.0	82.0	47.7	62.9	79.0	13.4

Table D.1: Probability of successful phase transition

Cost of Clinical Trial

The cost estimates of the clinical trials are obtained from [251] and [215]. The estimates used in simulation are included in Table D.2.

Therapeutic Area	Cost ($ millions)					
	Preclinical to Phase 1	Phase 1 to Phase 2	Phase 2 to Phase 3	Phase 3 to NDA	NDA to Approval	Preclinical to Approval
Oncology	3.9	9.9	24.7	34.7	2.9	76.0
Metabolic/Endocrinology	3.9	3.1	26.6	45.4	2.9	81.9
Cardiovascular	3.9	4.8	15.4	84.4	2.9	111.4
CNS	3.9	8.6	30.6	31.6	2.9	77.5
Autoimmune/Inflammation	3.9	8.8	36.3	35.7	2.9	87.6
Genitourinary	3.9	6.8	32.1	19.3	2.9	65.0
Infectious Disease	3.9	9.2	31.3	39.5	2.9	86.8
Ophthalmology	3.9	11.7	30.4	34.9	2.9	83.7

Table D.2: Cost of a clinical trial

Duration of Clinical Trial

The duration of clinical trials are obtain from [288] and [287]. The estimates used in simulation are included in Table D.3.

Therapeutic Area	Duration (months)					
	Preclinical to Phase 1	Phase 1 to Phase 2	Phase 2 to Phase 3	Phase 3 to NDA	NDA to Approval	Preclinical to Approval
Oncology	12.0	36.4	41.0	49.9	10.0	149.3
Metabolic/Endocrinology	12.0	10.8	31.5	32.5	10.0	96.9
Cardiovascular	12.0	12.6	34.2	40.3	10.0	109.1
CNS	12.0	11.1	31.1	34.5	10.0	98.7
Autoimmune/Inflammation	12.0	11.2	32.7	32.6	10.0	98.5
Genitourinary	12.0	12.6	26.2	33.5	10.0	94.3
Infectious Disease	12.0	18.7	31.7	35.6	10.0	108.0
Ophthalmology	12.0	18.2	27.4	34.3	10.0	101.9

Table D.3: Duration of clinical trials

Market Value of Drugs

One of the important parameter of simulation is the market value of the drugs. The estimates of the market value are obtained from the various market research reports.

From Approval values of drugs, the value of drugs intermediate phases are obtained using the Equation 5.1. The estimates of the market value used in simulation are included in Table D.4.

	Value ($ millions)					
Therapeutic Area	Preclinical	Phase 1	Phase 2	Phase 3	NDA	Approval
Oncology	0.3	8.1	55.3	460.7	2020.1	2669.5
Metabolic/Endocrinology	3.2	13.9	32.5	198.3	531.1	685.0
Cardiovascular	3.3	14.1	37.3	207.3	789.3	1016.9
CNS	3.8	14.9	45.7	307.1	890.3	1153.0
Autoimmune/Inflammation	1.1	9.7	33.5	274.4	683.4	893.7
Genitourinary	2.0	11.5	34.9	206.2	457.8	600.6
Infectious Disease	0.8	9.1	38.6	207.8	543.0	681.0
Ophthalmology	1.8	11.0	38.8	209.7	544.0	750.0

Table D.4: Market value of drugs

Correlation Between Compounds

The drugs across therapeutic areas share similar features because of which the outcomes of the clinical trials across for the drugs are correlated. We use the block correlation matrix structure for modelling outcome for different clinical trial projects as presented in the Figure D-1. We assume a correlation of 0.05 for the clinical trial projects across different therapeutic areas and 0.20 for clinical trials in same therapeutic areas.

Royalty Asset

As mentioned earlier, the cash flows from royalty assets are estimated using a discounted cash flow analysis. The details of the analysis are included in the Figure D-2. We modelled the cashflows for the royalty asset for 10 years[1]. We assume an initial year 1 sale of 100 million$ that increases to 1000 million$ in period of 5 years and it grows at a constant rate of 2.5%. The royalty is acquired for 5% of the total sales. We find that net present value of the cash flows is 222million$. We discount it using

[1]The patent for drugs usually lasts for 10 years

	ONCO	ONCO	ONCO	META	META	META	CARD	CARD	CARD	CNS	CNS	CNS	AUTO	AUTO	AUTO	GENI	GENI	GENI	INFE	INFE	INFE	OPHT	OPHT	OPHT
ONCO	1.00	0.20	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
ONCO	0.20	1.00	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
ONCO	0.20	0.20	1.00	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
META	0.05	0.05	0.05	1.00	0.20	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
META	0.05	0.05	0.05	0.20	1.00	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
META	0.05	0.05	0.05	0.20	0.20	1.00	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CARD	0.05	0.05	0.05	0.05	0.05	0.05	1.00	0.20	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CARD	0.05	0.05	0.05	0.05	0.05	0.05	0.20	1.00	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CARD	0.05	0.05	0.05	0.05	0.05	0.05	0.20	0.20	1.00	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CNS	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	1.00	0.20	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CNS	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	1.00	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
CNS	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	0.20	1.00	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
AUTO	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	1.00	0.20	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
AUTO	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	1.00	0.20	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
AUTO	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	0.20	1.00	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
GENI	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	1.00	0.20	0.20	0.05	0.05	0.05	0.05	0.05	0.05
GENI	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	1.00	0.20	0.05	0.05	0.05	0.05	0.05	0.05
GENI	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	0.20	1.00	0.05	0.05	0.05	0.05	0.05	0.05
INFE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	1.00	0.20	0.20	0.05	0.05	0.05
INFE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	1.00	0.20	0.05	0.05	0.05
INFE	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	0.20	1.00	0.05	0.05	0.05
OPHT	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	1.00	0.20	0.20
OPHT	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	1.00	0.20
OPHT	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.20	0.20	1.00

Figure D-1: Block correlation matrix

10% rate[2] for the associated risk in the cash flows and find that value of a royalty asset is 200 million $.

$ millions	0	1	2	3	4	5	6	7	8	9	10
Sales		100	325	550	775	1,000	1,025	1,051	1,077	1,104	1,131
Royalty		5	16	28	39	50	51	53	54	55	57
% of sales		*5.0%*	*5.0%*	*5.0%*	*5.0%*	*5.0%*	*5.0%*	*5.0%*	*5.0%*	*5.0%*	*5.0%*
Discounted cash flow		5	13	21	26	31	29	27	25	23	22
Discount period		*1.0*	*2.0*	*3.0*	*4.0*	*5.0*	*6.0*	*7.0*	*8.0*	*9.0*	*10.0*
Discount rate		*10.0%*	*10.0%*	*10.0%*	*10.0%*	*10.0%*	*10.0%*	*10.0%*	*10.0%*	*10.0%*	*10.0%*
Discount factor		*0.91*	*0.83*	*0.75*	*0.68*	*0.62*	*0.56*	*0.51*	*0.47*	*0.42*	*0.39*
Market value	200										
Net present value	222										
Discount	10.0%										

Figure D-2: Discounted Cash Flow analysis of royalty investement

D.2 Sensitivity Analsysis

Varying Leverage Ratio

We changed the leverage ratio of portfolio from 80:20 (debt:equity) to 60:40, 40:60, 20:80 and 0:100 and estimated the portfolio performance. The performance of the portfolio is included in Table D.5.

[2]https://www.nytimes.com/2017/07/08/business/dealbook/drug-prices-private-equity.html

Debt to Equity Ratio	Benchmark 80:20	60:40	40:60	20:80	0:100	Optimized 80:20	60:40	40:60	20:80	0:100
Capital Structure										
Total ($ millions)	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800
Debt ($ millions)	10,240	7,680	5,120	2,560	0	10,240	7,680	5,120	2,560	0
Equity ($ millions)	2,560	5,120	7,680	10,240	12,800	2,560	5,120	7,680	10,240	12,800
Tenor (years)	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5
Debt with 2% coupon										
Prob. of Default (bp)	19.0	0.2	0.0	0.0	0.0	18.6	0.0	0.0	0.0	0.0
E[Loss] (bp)	1.2	0.0	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0
Equity										
E[Cum. ROE] (%)	499.9	267.6	186.1	144.7	119.9	472.5	252.7	176.1	137.2	113.9
SD[Cum. ROE] (%)	306.3	156.3	104.5	78.4	62.7	273.4	138.7	92.7	69.5	55.6
Ann. Sharpe Ratio	0.42	0.43	0.44	0.44	0.45	0.44	0.46	0.46	0.47	0.48
Geometric Ann. E[Cum. ROE] (%)	13.2	9.4	7.5	6.4	5.6	12.8	9.1	7.3	6.1	5.4
Arithmetic E[Ann. ROE] (%)	34.5	18.5	12.8	10.0	8.3	32.6	17.5	12.2	9.5	7.9
Arithmetic SD[Ann. ROE] (%)	21.0	10.7	7.2	5.4	4.3	18.8	9.5	6.4	4.8	3.8
Geometric E[Ann. ROE] (%)	11.8	8.7	7.0	6.0	5.3	11.6	8.5	6.8	5.8	5.2
Geometric SD[Ann. ROE] (%)	6.8	3.6	2.8	2.4	2.1	6.6	3.2	2.6	2.2	1.9
Prob. of Wipeout (%)	0.2	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0
Prob. of Loss (%)	1.9	1.7	1.2	0.9	0.6	1.8	1.4	1.0	0.7	0.5
Prob. of Ann. ROE > 10% (%)	72.1	38.3	14.1	4.1	1.0	71.6	34.3	10.3	2.1	0.3

(a) Bechmark and optimized portfolio performance varying leverage ratio; debt: equity (80:20, 60:40, 40:60, 20:80 and 0:100)

Debt to Equity Ratio	Benchmark 80:20	60:40	40:60	20:80	0:100	Optimized 80:20	60:40	40:60	20:80	0:100
Expected Approvals										
Oncology	1.1	1.5	1.5	1.5	1.5	0.6	0.8	0.8	0.8	0.8
Metabolic/Endocrinology	3.3	3.5	3.5	3.5	3.5	5.0	5.4	5.4	5.4	5.4
Cardiovascular	2.0	2.4	2.5	2.5	2.5	2.1	2.7	2.8	2.8	2.8
CNS	1.7	2.2	2.3	2.3	2.3	1.0	1.2	1.3	1.3	1.3
Autoimmune/Inflammation	2.6	3.0	3.0	3.0	3.0	1.5	1.8	1.9	1.9	1.9
Genitourinary	3.8	3.9	3.9	3.9	3.9	4.5	4.6	4.6	4.6	4.6
Infectious Disease	3.4	3.6	3.6	3.6	3.6	4.4	4.8	4.9	4.9	4.9
Ophthalmology	2.8	3.1	3.1	3.1	3.1	3.2	3.6	3.6	3.6	3.6
Total	20.8	23.1	23.5	23.5	23.5	22.4	24.9	25.3	25.3	25.3

(b) Number of approved drugs for varying leveraged portfolio.

Table D.5

Varying Probability of Success

We vary the probability of the success to observe the effect on the performance. The PoS values used in the study are baseline PoS values obtained on the basis of historical data and literature. We hypothesize that experts involved in portfolio creation can select the drug development projects that can perform better than baseline values. We simulated all equity financed megafund varying the PoS from given in Table D.1 to 1.05, 1.10,1.15, 1.20 and 1.25 times the values in Table D.1. The detailed results are included in Table D.6.

	Benchmark						Optimized					
p=prob of phase transition	1.00p	1.05p	1.10p	1.15p	1.20p	1.25p	1.00p	1.05p	1.10p	1.15p	1.20p	1.25p
Capital Structure												
Total ($ millions)	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800
Debt ($ millions)	0	0	0	0	0	0	0	0	0	0	0	0
Equity ($ millions)	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800	12,800
Tenor (years)	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5	14.5
Equity												
E[Cum. ROE] (%)	119.9	152.3	189.9	233.6	281.3	324.7	113.9	144.8	180.6	222.3	266.8	307.6
SD[Cum. ROE] (%)	62.7	70.3	78.1	86.6	95.2	101.9	55.6	62.0	68.4	75.2	81.7	86.5
Ann. Sharpe Ratio	0.45	0.52	0.60	0.67	0.74	0.81	0.48	0.56	0.65	0.73	0.82	0.90
Geometric Ann. E[Cum. ROE] (%	5.6	6.6	7.6	8.7	9.7	10.5	5.4	6.4	7.4	8.4	9.4	10.2
Arithmetic E[Ann. ROE] (%)	8.3	10.5	13.1	16.1	19.4	22.4	7.9	10.0	12.5	15.3	18.4	21.2
Arithmetic SD[Ann. ROE] (%)	4.3	4.8	5.4	6.0	6.6	7.0	3.8	4.3	4.7	5.2	5.6	6.0
Geometric E[Ann. ROE] (%)	5.3	6.3	7.4	8.4	9.4	10.3	5.2	6.1	7.2	8.2	9.2	10.0
Geometric SD[Ann. ROE] (%)	2.1	2.1	2.0	2.0	1.9	1.9	1.9	1.9	1.8	1.8	1.7	1.7
Prob. of Wipeout (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Prob. of Loss (%)	0.6	0.2	0.0	0.0	0.0	0.0	0.5	0.1	0.0	0.0	0.0	0.0
Prob. of Ann. ROE > 10% (%)	1.0	3.3	9.2	21.8	40.2	57.8	0.3	1.5	5.4	15.9	33.9	52.8

(a)

	Benchmark						Optimized					
p=prob of phase transition	1.00p	1.05p	1.10p	1.15p	1.20p	1.25p	1.00p	1.05p	1.10p	1.15p	1.20p	1.25p
Expected Approvals												
Oncology	1.5	1.9	2.2	2.6	3.1	3.5	0.8	1.0	1.2	1.4	1.6	1.8
Metabolic/Endocrinology	3.5	4.2	5.0	5.9	6.9	7.8	5.4	6.5	7.8	9.3	10.9	12.4
Cardiovascular	2.5	3.0	3.6	4.2	4.9	5.5	2.8	3.3	3.9	4.6	5.4	6.0
CNS	2.3	2.7	3.2	3.8	4.5	5.0	1.3	1.5	1.8	2.1	2.5	2.8
Autoimmune/Inflammation	3.0	3.6	4.3	5.1	6.1	6.8	1.9	2.2	2.7	3.2	3.7	4.2
Genitourinary	3.9	4.7	5.6	6.6	7.8	8.8	4.6	5.5	6.6	7.8	9.2	10.3
Infectious Disease	3.6	4.4	5.2	6.2	7.1	8.0	4.9	5.9	7.0	8.3	9.3	10.4
Ophthalmology	3.1	3.7	4.5	5.3	6.3	7.3	3.6	4.4	5.4	6.5	7.8	9.1
Total	23.5	28.2	33.6	39.9	46.7	52.9	25.3	30.4	36.3	43.1	50.3	56.9

(b)

Table D.6: Portfolio performance varying Probability of Success

Varying Market Value of Drugs

We modified the value of approved market drugs from D.4 to two different set of values: moderate (minor modification in the market value) and aggressive (large modifications in the market value) given in the Table D.7. The megafund performance varying the market value of drugs is included in the Figure D.8.

Therapeutic Area	Value ($ millions) Preclinical	Phase 1	Phase 2	Phase 3	NDA	App
Oncology	0.0	6.0	55.0	529.3	2,755.8	3,640.2
Metabolic/Endocrinology	3.2	13.9	32.5	198.3	531.1	685.0
Cardiovascular	1.0	9.7	29.3	186.6	789.3	1,016.9
CNS	1.2	10.0	36.8	286.4	890.3	1,153.0
Autoimmune/Inflammation	0.0	7.3	29.7	274.4	683.4	893.7
Genitourinary	2.0	11.5	34.9	206.2	457.8	600.6
Infectious Disease	0.8	9.1	38.6	207.8	543.0	681.0
Ophthalmology	1.8	11.0	38.8	209.7	544.0	750.0
Vaccines	0.0	0.0	0.0	0.0	0.0	0.0

(a) Moderate Market Values

Therapeutic Area	Value ($ millions) Preclinical	Phase 1	Phase 2	Phase 3	NDA	App
Oncology	0.0	7.9	69.1	623.4	3,215.6	4,246.9
Metabolic/Endocrinology	0.6	8.8	22.7	165.6	459.9	593.7
Cardiovascular	0.8	9.4	29.3	186.6	789.3	1,016.9
CNS	0.9	9.7	36.8	286.4	890.3	1,153.0
Autoimmune/Inflammation	0.6	9.0	33.4	290.0	717.7	938.4
Genitourinary	0.0	7.2	26.7	181.0	406.6	533.9
Infectious Disease	0.0	6.5	33.1	191.2	506.6	635.6
Ophthalmology	0.0	7.1	32.0	189.2	498.4	687.5
Vaccines	0.0	0.0	0.0	0.0	0.0	0.0

(b) Aggressive Market Values

Table D.7: Market Value of Drugs

	Moderate		Aggressive	
	Benchmark	Optimized	Benchmark	Optimized
Capital Structure				
Total ($ millions)	12,900	13,200	13,100	13,300
Senior Debt ($ millions)	9,030	9,240	9,170	9,310
Junior Debt ($ millions)	1,290	1,320	1,310	1,330
Equity ($ millions)	2,580	2,640	2,620	2,660
Tenor (years)	14.5	14.5	14.5	14.5
Senior				
Prob. of Default (bp)	1.2	1.4	1.0	2.0
E[Loss] (bp)	0.0	0.1	0.1	0.1
Junior				
Prob. of Default (bp)	14.8	19.2	17.2	16.4
E[Loss] (bp)	5.7	6.9	5.6	6.6
Equity				
E[Cum. ROE] (%)	559.0	502.2	566.5	511.3
SD[Cum. ROE] (%)	350.2	294.7	361.1	308.2
Ann. Sharpe Ratio	0.41	0.44	0.40	0.43
Geometric Ann. E[Cum. ROE] (%	13.9	13.2	14.0	13.3
Arithmetic E[Ann. ROE] (%)	38.6	34.7	39.1	35.3
Arithmetic SD[Ann. ROE] (%)	24.1	20.2	24.8	21.2
Geometric E[Ann. ROE] (%)	12.5	11.9	12.6	12.0
Geometric SD[Ann. ROE] (%)	6.5	6.7	6.9	6.6
Prob. of Wipeout (%)	0.2	0.2	0.2	0.2
Prob. of Loss (%)	2.1	1.9	2.1	2.0
Prob. of Ann. ROE > 10% (%)	75.4	73.6	75.3	73.3

(a)

	Moderate		Aggressive	
	Benchmark	Optimized	Benchmark	Optimized
Expected Approvals				
Oncology	1.1	0.6	1.0	0.6
Metabolic/Endocrinology	3.1	4.8	3.4	4.8
Cardiovascular	2.1	2.6	2.1	2.9
CNS	1.5	1.1	1.6	1.3
Autoimmune/Inflammation	2.3	1.5	2.2	1.7
Genitourinary	3.5	3.9	3.8	3.0
Infectious Disease	3.3	5.1	3.6	5.9
Ophthalmology	2.7	2.0	3.1	2.3
Total	19.6	21.6	20.8	22.5

(b)

Table D.8: Portfolio performance on varying market value of drugs.

Appendix E

Finance for Impact: Quantifying the Returns of ESG Investing

E.1 Rank Autocorrelation

In Table E.1, we present the cross-sectional rank autocorrelation between vendor ESG scores and aggregate scores averaged over time. We find there is a 99% autocorrelation computed with a delay of one month, implying that ESG scores do not change significantly over short periods. However, lower values of rank autocorrelation at longer delay windows implies that scores change significantly over longer time windows. This pattern is consistent across regions. Hence, to rebalance our ESG portfolios, we use a time window of 12 months.

E.2 Implied Realized Alphas

In Table E.2, we present the realized and implied alphas of the individual portfolios using the Fama-French five-factor model. There is a high degree of consistency between the realized and implied alphas, hence validating the model described in Section 6.2.1.

	USA				Europe				Japan			
Delay (Months)	1	3	6	12	1	3	6	12	1	3	6	12
ISS: ESG	99	98	97	94	99	98	97	95	99	98	97	95
MSCI: ESG	99	97	95	90	99	97	94	90	99	97	94	89
Reprisk: ESG	98	96	93	88	98	96	93	86	98	94	89	79
SPGlobal: ESG	99	98	96	92	99	98	97	94	99	99	98	96
TVL: ESG	98	94	86	73	99	95	88	75	99	94	85	69
Moody's: ESG	99	98	96	94	99	98	97	95	99	98	96	94
AVG: ESG	99	98	96	92	99	98	97	94	99	98	96	92
PCA: ESG	99	99	98	97	99	99	98	97	99	99	98	97
MAHA: ESG	98	96	93	86	98	96	92	86	98	95	91	82
AVG_V: ESG	99	98	96	93	99	98	97	94	99	98	96	92
STV_V: ESG	95	88	82	71	95	90	83	74	96	91	85	76

Table E.1: Cross-sectional rank autocorrelation between vendor ESG scores and aggregate scores.

E.3 Sharpe Ratios of Combined Market and Treynor-Black ESG Portfolios

ESG-Data	Q-1 (R)	Q-1 (I)	Q-2 (R)	Q-2 (I)	Q-3 (R)	Q-3 (I)	Q-4 (R)	Q-4 (I)
US								
MSCI	-0.9 (1.4)	-0.0 (0.5)	1.9 (1.2)	0.7 (0.5)	1.8 (1.1)	1.3 (0.5)	1.3 (1.1)	2.1 (0.5)
SPGlobal	2.1 (1.3)	1.1 (0.5)	-0.0 (1.2)	1.0 (0.5)	0.8 (1.1)	1.0 (0.5)	1.3 (1.1)	1.0 (0.5)
ISS	0.5 (1.3)	0.5 (0.5)	0.3 (1.3)	0.9 (0.5)	1.9 (1.2)	1.1 (0.5)	1.4 (1.1)	1.5 (0.5)
Moody's	1.2 (1.4)	0.8 (0.5)	1.1 (1.2)	1.0 (0.5)	0.2 (1.1)	1.1 (0.5)	1.5 (1.2)	1.3 (0.5)
Reprisk	0.2 (1.1)	-0.4 (0.5)	-0.3 (1.2)	0.7 (0.5)	2.0 (1.3)	1.4 (0.5)	2.1 (1.3)	2.4 (0.5)
TVL	-0.3 (1.2)	0.7 (0.5)	1.8 (1.1)	0.9 (0.5)	1.0 (1.2)	1.1 (0.5)	1.6 (1.4)	1.3 (0.5)
Europe								
MSCI	7.5 (2.5)	7.2 (0.6)	7.7 (2.2)	8.1 (0.6)	9.7 (2.3)	8.7 (0.6)	9.6 (2.1)	9.6 (0.6)
SPGlobal	7.8 (2.3)	7.1 (0.6)	7.7 (2.4)	8.1 (0.6)	8.8 (2.3)	8.7 (0.6)	10.0 (2.2)	9.7 (0.6)
ISS	6.6 (2.4)	6.7 (0.6)	9.0 (2.4)	8.0 (0.6)	8.9 (2.3)	8.8 (0.6)	9.8 (2.1)	10.1 (0.6)
Moody's	7.7 (2.3)	7.2 (0.6)	8.2 (2.4)	8.1 (0.6)	9.1 (2.2)	8.7 (0.6)	9.3 (2.2)	9.6 (0.6)
Reprisk	9.0 (2.3)	9.0 (0.6)	9.9 (2.3)	8.6 (0.6)	6.8 (2.3)	8.2 (0.6)	8.5 (2.3)	7.8 (0.6)
TVL	7.8 (2.3)	8.2 (0.6)	8.6 (2.2)	8.3 (0.6)	9.8 (2.3)	8.5 (0.6)	8.3 (2.4)	8.7 (0.6)
Japan								
MSCI	6.9 (3.2)	7.4 (0.5)	7.8 (3.1)	7.9 (0.5)	9.2 (3.2)	8.3 (0.5)	8.4 (3.1)	8.8 (0.5)
SPGlobal	6.4 (3.2)	6.2 (0.5)	7.2 (3.2)	7.6 (0.5)	8.3 (3.2)	8.6 (0.5)	10.4 (3.1)	10.0 (0.5)
ISS	6.5 (3.2)	6.6 (0.5)	7.1 (3.2)	7.7 (0.5)	8.6 (3.2)	8.5 (0.5)	10.2 (3.1)	9.6 (0.5)
Moody's	6.8 (3.2)	6.9 (0.5)	8.1 (3.3)	7.8 (0.5)	8.7 (3.2)	8.4 (0.5)	8.7 (3.1)	9.3 (0.5)
Reprisk	7.5 (3.3)	7.5 (0.5)	8.7 (3.2)	7.9 (0.5)	7.0 (3.1)	8.2 (0.5)	9.2 (3.1)	8.6 (0.5)
TVL	9.0 (3.1)	8.1 (0.5)	6.6 (3.2)	8.1 (0.5)	8.0 (3.1)	8.1 (0.5)	8.8 (3.2)	8.0 (0.5)

Table E.2: Model-implied and Realized Alphas
Average annual model-implied and realized alphas of quantile ESG portfolios with respect to Fama-French five-factor regressions, based on data from 2014 to 2020. Portfolios are rebalanced once a year.

	USA			Europe			Japan		
Portfolio	±40%	±25%	±10%	±40%	±25%	±10%	±40%	±25%	±10%
Mkt	0.75			0.05			0.37		
ISS: ESG	1.07	1.03	1.01	0.20	0.18	0.18	**0.64**	**0.65**	0.66
MSCI: ESG	1.03	1.08	1.10	**0.33**	**0.36**	0.32	0.54	0.58	0.34
Reprisk: ESG	0.81	0.79	0.87	0.22	0.31	**0.53**	0.45	0.48	0.52
SPGlobal: ESG	0.68	0.66	0.76	0.01	0.03	0.16	0.49	0.51	0.62
TVL: ESG	0.80	0.75	0.54	0.14	0.18	0.06	0.23	0.24	0.33
Moody's: ESG	0.87	0.91	0.95	-0.03	-0.01	-0.05	0.38	0.36	0.55
INDI-AVG: ESG	0.91	0.91	0.94	0.14	0.18	0.21	0.47	0.49	0.56
AVG: ESG	**1.13**	**1.25**	1.13	0.24	0.26	0.28	0.59	0.58	0.50
PCA: ESG	0.94	0.98	0.94	0.09	0.12	0.18	0.51	0.52	0.51
MAHA: ESG	1.12	1.15	**1.14**	0.26	0.31	0.30	0.62	0.62	0.57
AVG_{vote}: ESG	1.09	1.16	1.10	0.24	0.28	0.33	0.58	0.59	0.51
STV_{vote}: ESG	0.84	0.79	0.63	0.14	0.20	0.46	0.55	0.57	**0.72**
OPT: ESG	1.21	1.26	1.25	0.25	0.29	0.52	0.64	0.65	0.53

Table E.3: performance of Market+ Treynor Black ESG Portfolios.
Sharpe ratios of combined market and ESG Treynor-Black portfolios. The market portfolios used for different regions are the MSCI USA, MSCI Europe and MSCI Japan indices. The weight of the market index is fixed at 0.5. The highest realizable Sharpe ratio in each portfolio is given in bold.

Appendix F

Behavioural Finance: Real-time extended psychophysiological analysis of financial risk processing

F.1 Market Efficiency

Efficient Market Hypothesis It states that security prices reflect all the available information. It is based on the standard economic paradigm of rational expectations postulates that all financial decision makers are rational agents who maximize their expected utility functions under all market circumstances [255]. However, this paradigm does not consider the role of psychophysiological state in financial risk processing of traders. In recent decades, the rational expectations paradigm has been challenged by empirical observations in the behavioral finance literature.

Adaptive Market Hypothesis It combines the principles of the efficient market hypothesis with behavioral finance to explain the deviations from the market efficiency [179]. It points out that investors are not always rational and security prices do not always trade at the fair value. It incorporates the psychological and behavioral characteristics of traders to explain deviations from rationality.

F.2 Physiological Data Collection

Subjects: Fifty-five professional traders working at a global financial institution participated in this study. The traders were divided into seven groups according to their trading desks. Six groups were based in the institution's headquarters, and one in a regional office. The traders were recruited by the financial institution and varied in age, gender, trading experience, and types of financial product they traded. Each group of traders typically belonged to the same trading desk and specialized in a particular type of financial products, such as municipal bonds, mortgage-backed securities, equities, and power and natural gas.

Physiology Data Collection Procedure The physiology data collection required seven weeks, and took place between March 25, 2019, and June 21, 2019. Each group of traders recorded their physiological waveforms via the Empatica E4 wristbands over five consecutive workdays (with a few exceptions). On the Friday before each week of data collection, the study team met with the traders who would participate in the following week to explain the data collection procedure and obtain their informed consent. Each trader received written instructions to wear the Empatica E4 wristband (as shown in Figure F-1) on the wrist of their non-dominant side from the time they arrived at the office until the end of the workday, Monday to Friday of their week. The collected physiological waveforms were downloaded to a secure device by the study team at the end of Monday, Wednesday, and Friday during each week. To ensure the anonymity of the traders, each trader was assigned a participant ID in the format of "[wristband serial #]_[week #]".

F.3 Market and Financial Data Collection

Market Time Series Data: To understand how real-time fluctuations of different market sectors affect trader's psychophysiological activation, we acquired data for 10 commonly used market indices and futures prices from three sources, as shown in Table F.2. The market time series data was recorded during the same period as the physiology data collection (between March 25, 2019, and June 21, 2019). We defined the average of bid and ask prices as the futures prices.

Financial Trading Data: To understand how trading affects the levels of psychophysiological activation of each trader, we acquired the financial trading records of all traders who participated in our study from the global financial institution. The features of the financial data we used included trade price, quantity, notional amount, execution timestamp, and the lifecycle state of the trade, as well as the market sector and type of financial product being traded.

The number of transactions made by traders in each week is shown in Table F.3. We observe significant variation in the number of transactions across different weeks. In particular, traders in week 5 made many more transactions than the others. In our feedback session, we learned that most of the traders in week 5 used trading algorithms to place these transactions automatically.

F.4 Trader Survey

We asked traders to fill a survey during the trading days that included multiple questions. Some of them are years of trading experience, highest degree or level of school completed, gender, average number of transactions placed, business division they were part of.

F.5 Data Preprocessing

Preprocessing Physiological Waveforms: The physiological waveforms collected by the Empatica wristband (Table F.1) are measured in different frequencies. To reduce the measurement noise and align different physiologies in a multivariate time series with the same frequency, we first downsampled each raw physiological waveform to 1Hz by averaging the measurement values within each second. In our subsequent analysis, we removed the wrist accelerations (ACC) due to large irregularities which remained even after smoothing to 1Hz.

The inter-beat interval (IBI) measures the length of the time interval between two successive heartbeats, and does not have a regular frequency. Due to a measurement limitation of the Empatica E4 wristband, the IBI measurements for most traders were missing. To address this issue, we computed an average IBI time series using the instantaneous heart rate, HR_t: $IBI_t = 60/HR_t$.

Average IBI: Empatica E4 wristband needs 10 seconds of data for the heart rate calculation. Hence, IBI values obtained using instantanous heart rate is IBI averaged over 10 seconds. Hence, IBI_t values used in this analysis are average IBI values. When average IBI values are compared to the instantaneous IBI (when instantaneous IBI is available for longer durations), the mean squared error is 0.011 seconds and the error is considerably small as compared to the IBI values (0.6-1 second). In our analysis average IBI is one of the signal used for heart rate variability feature extraction that in turn was used for activation levels computation. We observed statistically significant relationships between activation values computed using average IBI and market, trading signals.

Remove Inactive Physiological Waveforms: We performed a visual inspection of each trader's physiological waveforms. We observed that 13 out of 55 traders had inactive periods in their physiology measurements when their BVP and EDA waveforms remained identically zero, as shown in Figure F-2 (A). Since 8 out of these 13 traders showed inactive periods after 5PM ET, we hypothesize that the inactive periods were due to traders not turning off their wristbands as they left work.

Since no meaningful physiology data was collected during these inactive periods, it was necessary to identify and remove them from our analysis. To achieve this, we inspected each trader's physiological waveforms, manually labeled the start and end time points of each inactive period, and removed these inactive periods from the subsequent analysis.

Aggregate Block transactions In the financial trading data, we observed many instances where multiple transactions, each recorded with a unique trade ID, occurred within one second. Since it is highly unlikely that these transactions were placed independently by the trader, we consulted the global financial institution and learned that these transactions in fact belonged to one multi-leg “block trade” placed by the trader, but were recorded as individual transactions in the trading data. However, no feature in the trading data indicated whether two individual transactions belonged to the same block trade.

To aggregate the individual transactions back to the original block transactions, we adopted a simple heuristic: If the trader placed two consecutive transactions within one second, we assumed that they belonged to the same block trade. The number of aggregated block transactions is illustrated in Table ?? for three representative traders. We also validated the results with the number of transactions self-reported by the traders in their survey responses (column 5 in Table ??).

In most cases (as shown in Table ??), the number of aggregated block transactions agrees well with the value self-reported by the traders in their survey responses. However, we should caution that in a few cases, the number of block transactions is significantly above the self-reported values (which may not be accurate). Overall, this heuristic produced a reasonable approximation to the original multi-leg block transactions.

Align Market Time Series The time series of the futures prices (Table F.1) were recorded at the timestamps when the transactions were executed in the market. As a result, these timestamps are irregular and do not form a time series with a uniform frequency. To perform a statistical analysis requiring an evenly spaced time series (such as a Granger causality test), we aligned the market time series, m_t, to minute-level m_t^{align}. We used minute-level data since the USD index was recorded at the beginning of each minute. We defined the aligned time series value m_t^{align} at minute t as the time series value observed closest to and no later than time t, i.e. $m_t^{align} := m_\tau$ where $\tau := max_s\{s : s \leq t \text{ and } m_s \text{ was observed}\}$. For notational simplicity, we use m_t to denote the market time series aligned to minute level in the subsequent analysis.

F.6 Activation Feature Extraction

Activation Proportion Once we obtained a time series of activation for each trader, we were interested in finding the regions of high activation. These regions of high activation are characterized by large deviations from the mean of the distribution. To find these regions, we used heuristics that are generally used for measuring deviations from mean, specifically choosing those that are dynamic and with a smaller number of tunable parameters.

- $arousal_t > mean(arousal)_{day} + \alpha * std(arousal)_{day}$: If the value of activation deviates from the mean value of arousal over the day, then it is considered an outlier. The value of α was set to 1.5 for periods of mild activation, and 3.0 for extreme activation.

- $arousal_t > mean(arousal)_{t:t-60min} + \alpha * std(arousal)_{t:t-60min}$: If the value of arousal deviates from the mean value of arousal calculated over a rolling window of 60 minutes, then it is considered an outlier. The value of α was set to 1.5 for periods of mild activation, and 3.0 for extreme activation.

The second heuristic is adaptive over time. Due to its adaptive nature, however, it may miss small outliers next to extreme outliers that may be captured by the first method. Hence, a period is considered high activation when it is flagged by either of the above methods. Finally, we computed the activation proportion for each trader, defined as the percentage of active trading time when the trader is at high activation. We calculated two variants of this proportion:

- Mild activation proportion: this is defined as the percentage of active trading time when the trader is under mild to high psychophysiological activation. For its calculation, the value of α was set to 1.5 in the outlier detection method. This metric captured periods of mild to high activation in traders.

- Extreme activation proportion: this is defined as the percentage of active trading time when the trader is under extremely high activation. For its calculation, the value of α was set to 3.0 in the outlier detection method. This metric captured periods of extremely high activation in traders.

Activation Length We analyzed the length of periods of continuous activation, which gave us information about the recovery time from periods of high activation, which in turn gave us some idea about the well being of the trader. To calculate the activation length, we followed the following procedure: after determining the periods of high activation, we calculated the length of each continuous period. If the periods of high activation were 5 minutes apart or more, they were considered different activation periods. For each trader on a given day, we calculated the median of the length of different activation periods to calculate the activation length.

Average Activation We calculate the mean value of activation over the day for a trader to obtain the average activation. The traders having higher activation proportion or length will have high average activation. Hence average activation captures the information from both the metrics activation proportion and length.

F.7 Traders Feedback Session

Initially we invited all 55 traders who participated in our study to the individual feedback sessions. Due to their busy schedule and limited availability, 14 of the 55 traders signed up for the feedback sessions. Each feedback session is conducted in a standardized format. For each trader, we created a report which summarized our PP activation analysis for this trader and shared the report with each trader the day before the meeting. The report positioned the PP activation level of the trader relative to the cohort of 55 traders in our study. It also included the market indices which have Granger-causal relation to the trader's activation levels. During the feedback session, we summarized the results and asked them the following questions:

- What causes your extreme activation levels during certain trading days (if there were any)?
- What market events and indices do you monitor on a regular basis?

- Are you interested in a follow-up study to analyze the relationship between your PP activation and trading performance (measured by PnL)?

F.8 Trader's Activation and Financial Product Type

The traders who participated in our study worked in different business divisions and traded a variety of financial products, including securitized markets (Securitized), credit, municipals (Muni), commodities, equities, foreign exchange and local markets (FXLM), and G10 rates. We investigate the relationship between financial products and trader activation measured by the regression coefficients.

We find that traders working in different business divisions had different average activation levels in the decreasing order: Commodities, FXLM, G10 Rates, Equities, Municipals, Credit and Securitized Markets. In table F.7, we present the regression coefficient of column business division compared to row business division along with the p-value and find the pairs of business divisions with statistically significant difference in their activation levels. For example, value $0.26_{0.04}$ in row *Securitized* and column *G10 R.* means that the difference between regression coefficients of G10 Rates product type and Securitized product type is 0.26 with p-value 0.04, which implies that G10 Rates traders had statistically significant higher activation than Securitized traders. The higher activation of FXLM and Equity traders can be attributed to high volatility in the markets that affects the PnL in real time while the G10 rate transactions were of higher amount (Table F.7) that can potentially lead to higher trader activation. We conclude that traders working in different business divisions have different profiles of PP activation.

Physiological Waveform	**Frequency**	**Unit**
Heart Rate (HR)	1	Number of beats
Blood Volume Pressure (BVP)	64	Millimeters of Mercury (mmHg)
Electrodermal Activity (EDA)	4	Microsiemens (μS)
Skin Temperature (TEMP)	4	Celsius (°C)
Wrist Accelerations (ACC)	32	$\frac{1}{32}g$
Inter-Beat Interval (ACC)	Irregular	Second(s)

Table F.1: Physiological waveforms measured by the Empatica E4 wristband for each trader.
The IBI measures the length of the time interval between two successive heartbeats, and does not have a regular frequency.

Data Source	**Market Data**
Chicago Mercantile Exchange	US Treasury Futures (2Y/5Y/10Y)
	E-mini S&P 500 Futures
	Crude Oil Futures
Global Financial Institution	US Dollar Index
	Credit-Default Swap Indices (HY and IG)
Chicago Board Options Exchange	VIX Futures

Table F.2: Market time series data acquired from different data sources.

Total	**Week 1**	**Week 2**	**Week 3**	**Week 4**	**Week 5**	**Week 6**	**Week 7**
51,732	145	308	305	1,577	45,246	3,523	628

Table F.3: Total number of transactions made by the traders in each week.
There is significant variation in the number of transactions across different weeks. Maximum transactions are concentrated in week 5 due to the use of automatic trading algorithms by traders.

Trader ID	Date	Number of Trade	Number of Block Trade	Trader's Survey Response
A01A22	April 8	180	52	20
	April 9	279	14	20
	April 12	305	13	20
A020A2	April 29	64	22	15-50
	April 30	80	41	15-50
	April 12	305	13	15-50
	May 1	43	15	15-50
	May 2	110	75	15-50
	May 3	69	24	15-50
A02187	April 29	778	235	200-500
	April 30	1157	330	200-500
	May 2	946	233	200-500

Table F.4: Aggregating Transactions.

Number of transactions before and after aggregating the block transactions for few representative traders on different days. The number of block transactions (column 4) is much closer to the number of transactions self-reported by the traders (column 5) than the raw data (column 3).

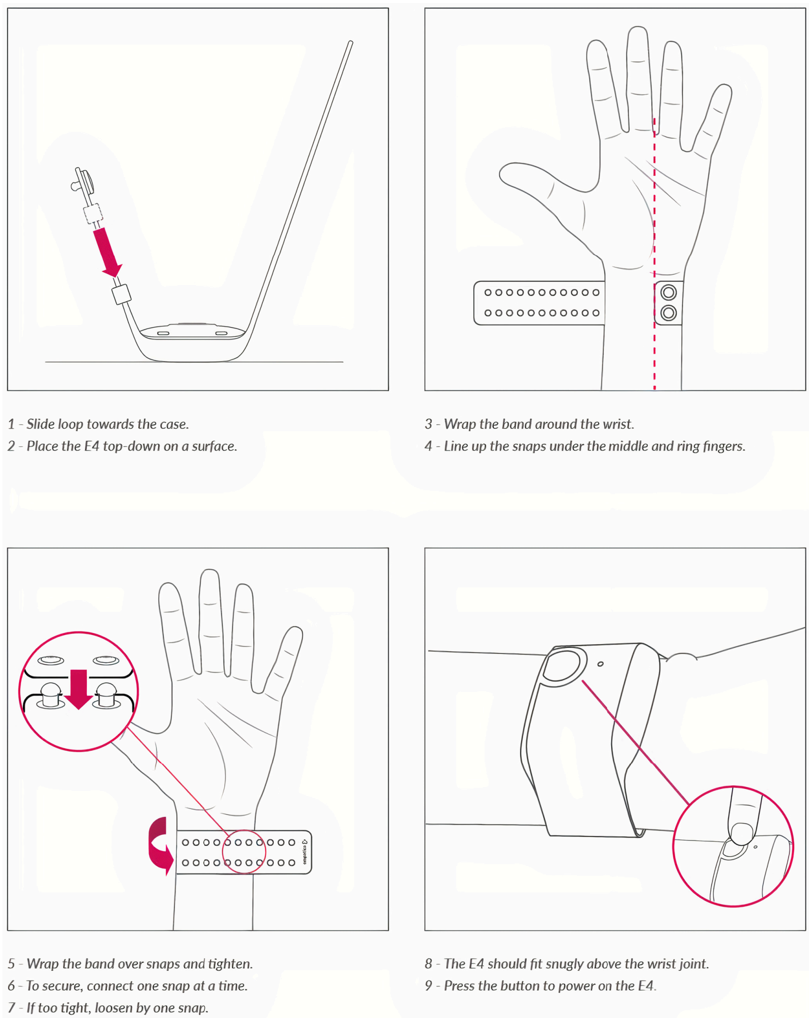


Figure F-1: E4 Wristband.
Placement of the Empatica E4 wristband which measured traders' physiological waveforms (Table F.1). This figure was shown to the traders to illustrate how to wear the Empatica E4 wristband on their non-dominant-side wrist.

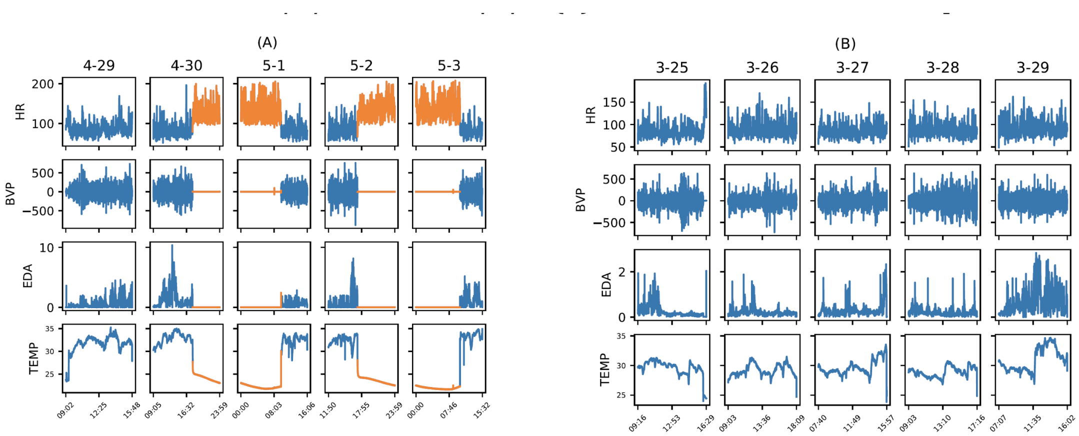


Figure F-2: Inactive Periods Detection.
Inactive periods in traders' physiological waveforms. (A) waveforms with inactive periods (plotted in orange). Based on the timestamps of the inactive periods, we suspect that this trader took off the wristband after work but did not terminate the measurements on 2019/4/30 and 2019/5/2. (B) waveforms with no inactive periods.

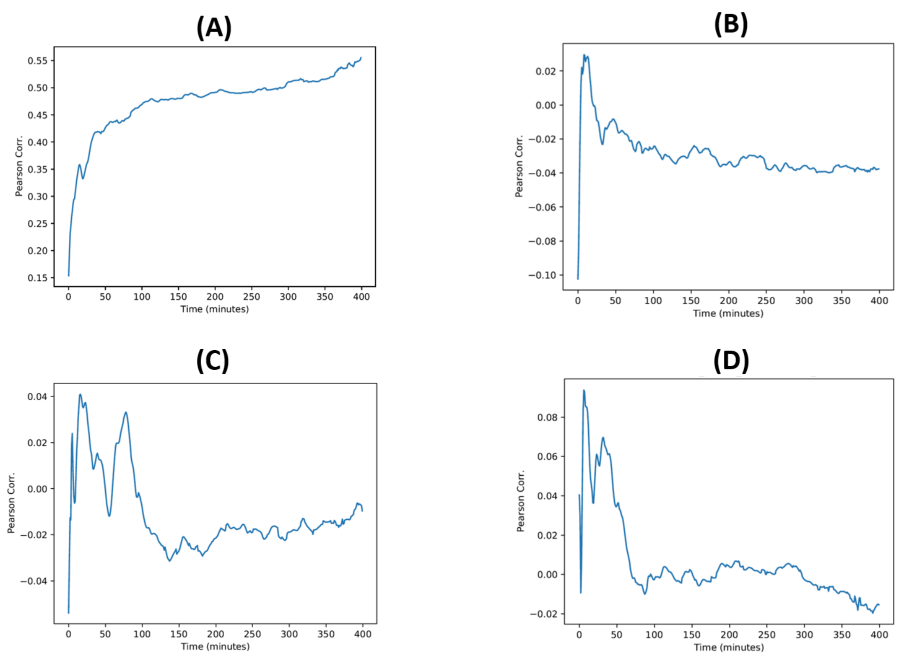


Figure F-3: Examples of rolling correlation between two physiology signals averaged over all traders. (A) EDA mean vs. EDA slope; (B) EDA mean vs.HR RMSSD; (C) EDA mean vs. HR mRRi; (D) EDA mean vs. BVP mean.

			Business									
		Experience	**Commodities**	**Credit**	**Equities**	**FXLM**	**G10 Rates**	**Municipals**	**Securitized Markets**	**Female**	**Amount**	**NumTrade**
	Experience	1.00(0.00)	0.24(0.00)	-0.27(0.00)	0.50(0.00)	-0.01(0.88)	-0.20(0.01)	-0.12(0.14)	-0.27(0.00)	-0.52(0.00)	-0.47(0.00)	0.62(0.00)
Business	**Commodities**	0.24(0.00)	1.00(0.00)	-0.15(0.08)	-0.21(0.01)	-0.15(0.07)	-0.19(0.02)	-0.11(0.20)	-0.15(0.07)	-0.25(0.00)	-0.36(0.00)	-0.17(0.04)
	Credit	-0.27(0.00)	-0.15(0.08)	1.00(0.00)	-0.20(0.01)	-0.14(0.08)	-0.18(0.03)	-0.10(0.23)	-0.14(0.08)	0.50(0.00)	0.07(0.41)	-0.17(0.04)
	Equities	0.50(0.00)	-0.21(0.01)	-0.20(0.01)	1.00(0.00)	-0.21(0.01)	-0.27(0.00)	-0.15(0.08)	-0.21(0.01)	-0.34(0.00)	-0.57(0.00)	0.72(0.00)
	FXLM	-0.01(0.88)	-0.15(0.07)	-0.14(0.08)	-0.21(0.01)	1.00(0.00)	-0.19(0.02)	-0.10(0.21)	-0.15(0.08)	0.21(0.01)	0.09(0.28)	-0.09(0.27)
	G10 Rates	-0.20(0.01)	-0.19(0.02)	-0.18(0.03)	-0.27(0.00)	-0.19(0.02)	1.00(0.00)	-0.13(0.11)	-0.19(0.02)	-0.01(0.94)	0.61(0.00)	-0.17(0.04)
	Municipals	-0.12(0.14)	-0.11(0.20)	-0.10(0.23)	-0.15(0.08)	-0.10(0.21)	-0.13(0.11)	1.00(0.00)	-0.10(0.21)	-0.05(0.55)	0.06(0.49)	-0.13(0.12)
	Securitized Markets	-0.27(0.00)	-0.15(0.07)	-0.14(0.08)	-0.21(0.01)	-0.15(0.08)	-0.19(0.02)	-0.10(0.21)	1.00(0.00)	0.03(0.73)	0.17(0.04)	-0.18(0.03)
	Female	-0.52(0.00)	-0.25(0.00)	0.50(0.00)	-0.34(0.00)	0.21(0.01)	-0.01(0.94)	-0.05(0.55)	0.03(0.73)	1.00(0.00)	0.19(0.02)	-0.26(0.00)
	Amount	-0.47(0.00)	-0.36(0.00)	0.07(0.41)	-0.57(0.00)	0.09(0.28)	0.61(0.00)	0.06(0.49)	0.17(0.04)	0.19(0.02)	1.00(0.00)	-0.44(0.00)
	NumTrade	0.62(0.00)	-0.17(0.04)	-0.17(0.04)	0.72(0.00)	-0.09(0.27)	-0.17(0.04)	-0.13(0.12)	-0.18(0.03)	-0.26(0.00)	-0.44(0.00)	1.00(0.00)

Table F.5: Correlation Between Features.

In this table, we present the correlation values between different features that were used for average activation regression (refer heading *Activation Level Attribution Regression* under section *Materials and Methods* in main text). The higher values are shaded with red colour and lower values are shaded with blue. The number of transactions is positively correlated with experience implying experience traders places more number of transactions but the transactions were smaller in volume. While, use of automated trading algorithms led to higher number of transactions for equity traders. We also observed high correlation between G10 rate traders and dollar volume that can potentially explain the high activation in G10 rates traders. The activation level attribution regression was done for 194 different trader-days and similarly the correlation table was calculated using 194 different data points due to unavailability of trader's survey data for some traders. The p-values of the correlation are included in the brackets. The high positive correlation values are shaded in red and the high negative correlation values are shaded in green.

	EDA mean	EDA slope	EDA events	BVP mean	BVP min	BVP max	HR mRRi	HR mHR	HR RMSSD	HR SDNN	TEMP slope
EDA mean	1.00 (0.00)	0.56 (0.00)	0.48 (0.00)	-0.01 (0.36)	0.02 (0.11)	-0.02 (0.16)	-0.02 (0.25)	0.01 (0.41)	-0.03 (0.03)	-0.01 (0.20)	0.01 (0.14)
EDA slope	0.56 (0.00)	1.00 (0.00)	0.87 (0.00)	0.15 (0.00)	-0.10 (0.00)	0.10 (0.00)	-0.10 (0.00)	0.09 (0.00)	-0.04 (0.00)	-0.01 (0.19)	-0.00 (0.99)
EDA events	0.48 (0.00)	0.87 (0.00)	1.00 (0.00)	0.19 (0.00)	-0.12 (0.00)	0.12 (0.00)	-0.12 (0.00)	0.10 (0.00)	-0.05 (0.00)	-0.01 (0.23)	0.00 (0.69)
BVP mean	-0.01 (0.36)	0.15 (0.00)	0.19 (0.00)	1.00 (0.00)	-0.52 (0.00)	0.53 (0.00)	-0.16 (0.00)	0.14 (0.00)	-0.01 (0.37)	0.05 (0.00)	-0.02 (0.05)
BVP min	0.02 (0.11)	-0.10 (0.00)	-0.12 (0.00)	-0.52 (0.00)	1.00 (0.00)	-0.75 (0.00)	0.07 (0.00)	-0.07 (0.00)	-0.04 (0.00)	-0.06 (0.00)	0.01 (0.07)
BVP max	-0.02 (0.16)	0.10 (0.00)	0.12 (0.00)	0.53 (0.00)	-0.75 (0.00)	1.00 (0.00)	-0.07 (0.00)	0.06 (0.00)	0.03 (0.01)	0.05 (0.00)	-0.02 (0.03)
HR mRRi	-0.02 (0.25)	-0.10 (0.00)	-0.12 (0.00)	-0.16 (0.00)	0.07 (0.00)	-0.07 (0.00)	1.00 (0.00)	-0.98 (0.00)	0.16 (0.00)	-0.11 (0.00)	0.05 (0.00)
HR mHR	0.01 (0.41)	0.09 (0.00)	0.10 (0.00)	0.14 (0.00)	-0.07 (0.00)	0.06 (0.00)	-0.98 (0.00)	1.00 (0.00)	-0.10 (0.00)	0.19 (0.00)	-0.05 (0.00)
HR RMSSD	-0.03 (0.03)	-0.04 (0.00)	-0.05 (0.00)	-0.01 (0.37)	-0.04 (0.00)	0.03 (0.01)	0.16 (0.00)	-0.10 (0.00)	1.00 (0.00)	0.68 (0.00)	-0.01 (0.39)
HR SDNN	-0.01 (0.20)	-0.01 (0.19)	-0.01 (0.23)	0.05 (0.00)	-0.06 (0.00)	0.05 (0.00)	-0.11 (0.00)	0.19 (0.00)	0.68 (0.00)	1.00 (0.00)	-0.03 (0.00)
TEMP slope	0.01 (0.14)	-0.00 (0.99)	0.00 (0.69)	-0.02 (0.05)	0.01 (0.07)	-0.02 (0.03)	0.05 (0.00)	-0.05 (0.00)	-0.01 (0.39)	-0.03 (0.00)	1.00 (0.00)

Table F.6: Correlation Between PP Signals.
Correlation matrix between physiology signals averaged over all traders. The p-values (shown in parenthesis) are associated with the one sample t-test with the null hypothesis that the average correlation is zero for a pair of physiology signals. The statistically significant correlations (p-value < 0.05) are color coded.

Business	**Credit**	**Muni.**	**Equity**	**G10**	**FXLM**	**Commodity**
Securitized	$0.00_{,0.99}$	$0.07_{,0.70}$	$0.21_{,0.26}$	$0.26_{,0.04}$	$0.30_{,0.13}$	$0.30_{,0.07}$
Credit		$0.07_{,0.70}$	$0.21_{,0.28}$	$0.25_{,0.06}$	$0.29_{,0.11}$	$0.30_{,0.10}$
Muni.			$0.14_{,0.48}$	$0.18_{,0.30}$	$0.22_{,0.14}$	$0.23_{,0.24}$
Equity				$0.04_{,0.81}$	$0.08_{,0.66}$	$0.09_{,0.58}$
G10 Rates					$0.04_{0.82}$	$0.05_{,0.78}$
FXLM						$0.01_{,0.96}$

Table F.7: Regression coefficient of trader activation in the column business division with row business division as the baseline and the associated p-value.

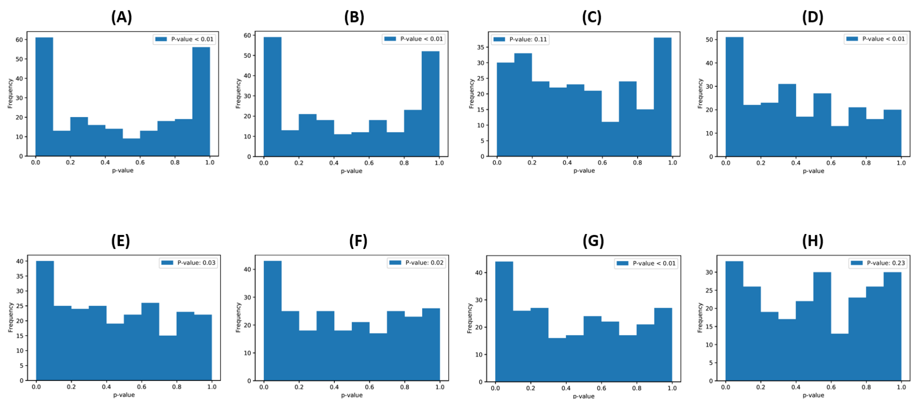


Figure F-4: Granger Causality Test Histograms

Histograms of p-values of Granger causality tests between each market index and overall activation for all traders (using a lag time of 10 minutes). For each histogram, we perform a one sample KS test with the null hypothesis that the p-values are uniformly distributed between 0 and 1 (i.e., no statistically significant Granger causality relations). (A) Credit Default Swap Index IG; (B) Credit Default Swap Index HY; (C) USD Index; (D) S&P 500 E-mini Futures Price; (E) 10Y US Treasury Futures Price; (F) 5Y US Treasury Futures Price; (G) Crude Oil Futures Price; (H) VIX Futures Price

Appendix G

xAI : Explainable Machine Learning Models of Consumer Credit Risk

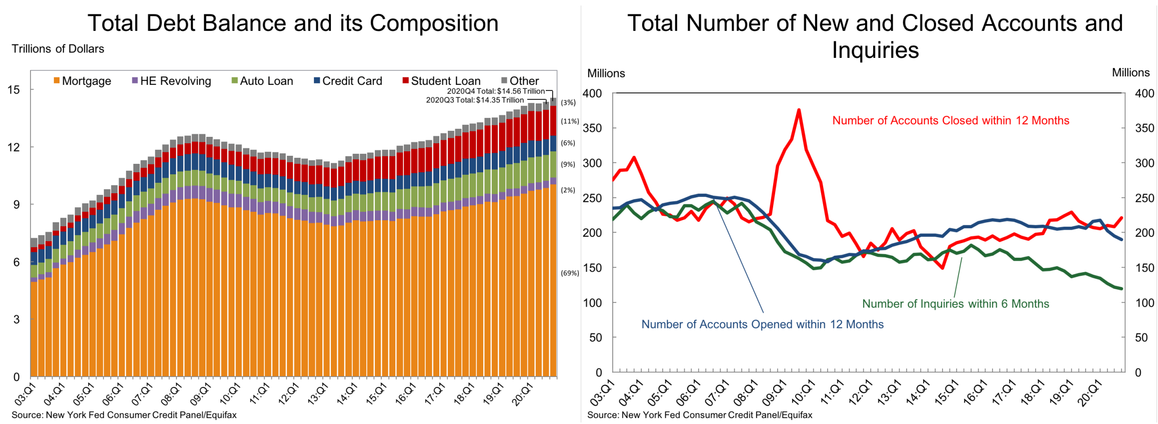
Figure G-1 Panel-a shows the total household debt balance in the United States and its composition over time. We observe that the household debt balance reached an all-time high at the end of the fourth quarter of 2020, with a value of \$14.56 trillion. The overall trend line of the debt balance is increasing. Figure G-1 Panel-b shows the total number of new accounts opened over time, along with the number of inquiries and number of accounts closed. The total number of accounts opened in the last single-year period measured is 189.6 million. This enormous debt balance and number of new accounts opened shows the importance of credit risk management. The data and the Figure G-1 were obtained from the New York Fed Consumer Credit Panel/Equifax.

G.1 Dataset Details

G.1.1 Glossary of Relevant Credit Modeling Terms

The following definitions provide helpful context for the description of the dataset's features.

1. Line of Credit: An agreement to provide credit.



(Panel-a) **(Panel-b)**

Figure G-1: Household Debt.
The plots were obtained from the quarterly report on household debt and credit released by the New York Fed.

2. Revolving Line of Credit: A line of credit with a maximum amount that the borrower can choose to use each month. The most common example is a credit card.

3. Installment Line of Credit: A line of credit with a fixed loan amount and a fixed monthly payment. A mortgage is a common example.

4. Delinquent: A line of credit is delinquent if its payments are not made in a timely manner.

5. Utilization: The amount still owed divided by the total amount borrowed; the fraction of available credit currently in use.

G.1.2 Explanation of Predictor Features

In addition to the binary target variable (Risk Classification), each credit applicant is characterized by 23 predictor features, 21 continuous and 2 categorical. These are:

1. A condensed version of the borrower's credit risk computed by FICO using all credit bureau information (ExternalRiskEstimate)

2. Months since the very first line of credit was established (MSinceFirstLOC)

3. Months since the newest line of credit was established (MSinceNewestLOC)
4. Average age in months of all existing lines of credit (AvgAgeOfLOC)
5. Number of lines of credit not currently delinquent (NumLOCNotDelq)
6. Number of lines of credit ever been 60 or more days delinquent (NumLOC60PlusDaysDelq)
7. Number of lines of credit ever been 90 or more days delinquent (NumLOC90PlusDaysDelq)
8. Percentage of lines of credit never been delinquent (PercentLOCNeverDelq)
9. Number of months since the most recent delinquency (MSinceMRecentDelq)
10. Maximum delinquency in days in the past year (MaxDelqLast12M)
11. Maximum delinquency ever in days (MaxDelqEver)
12. Total number of lines of credit established (NumTotalLOC)
13. Number of lines of credit established in the past year (NumLOCInLast12M)
14. Percentage of lines of credit that are installment lines of credit (PercentInstLOC)
15. Months since the newest request for a new line of credit excluding those requested in the past week (MSinceNewLOCReqExPastWeek)
16. Number of requests for new lines of credit in the last 6 months (NumLOCReqLast6M)
17. Number of requests for new lines of credit in the last 6 months excluding those requested in the past week (NumLOCReqLast6MExPastWeek)
18. Fraction of all revolving credit limits in use (FracRevLOCLimitUse)
19. Fraction of all installment lines of credit in use (FracInstLOCUse)
20. Number of revolving lines of credit with outstanding balances (NumRevLOCWBalance)

21. Number of installment lines of credit with outstanding balances (NumInst-LOCWBalance)

22. Number of bank loans and national loans (a subset of all revolving trades) with an outstanding balance of at least 75% of the credit limit (NumBank/NatlLoansWHighUtil)

23. Percentage of lines of credit with outstanding balances (PercentLOCWBalance)

G.1.3 Data Cleaning

Our dataset contains special values, negative integers that are interpreted symbolically and do not hold any numeric significance. As a result, we cannot directly feed them into our machine learning models. We either have to drop them or encode them appropriately. A large fraction of the data points in our HELOC dataset contains at least one special value (7,957 of the 10,459 data points). Hence, dropping all such data points is infeasible.

In addition, the dataset has 588 records that solely contain the special value -9 for all feature values. 331 of these data points are labeled as non-creditworthy, and 266 as creditworthy. This is a problem for any model because the same input vector will produce opposite target labels. This happens because a borrower will receive a special value if they are either a VIP and do not need to be investigated, or if they have no bureau record at all, i.e., they have no credit history. Such data points are dropped in our analysis.

We find special values are concentrated in 9 of our 23 input features. A standard technique for dealing with special values is to replace them with the mean values of the respective feature. A simple example illustrates that this is not a meaningful approach for our dataset. If a borrower has never had a delinquency, she will have the -7 (Condition not met) special value for the feature "months since most recent delinquency." Clearly, replacing the feature value with the mean is not correct, since she will be moved from a desirable value of the feature to a less desirable one. To handle these special values, we used binning techniques.

G.1.4 Data Pre-processing

The records in our dataset contain special values, which require a careful approach. We deal with special values by discretizing continuous features into bins. The advantage of binning is that special values can be treated as a separate bin, and any outliers can be consolidated. Once a binning schema has been decided, a feature can be represented using one-hot encoding and weight of evidence (WoE) encoding.

In one-hot encoding, a feature that contains n bins can be represented as an n-dimensional vector f. If a feature value belongs to bin_i then the value of its kth dimension $f_k = 1$ if $k = i$ and 0 otherwise, for $k \in [0, n)$. The drawbacks of one-hot encoding are that bins are treated as unordered categories, and sparsity is introduced. Sparse features can result in overfitting and biased parameters in a model if the training dataset is small. In addition, for continuous variables, there is no clear-cut formula to define the binning schema, and choosing it manually can be sub-optimal.

Weight of evidence (WoE) encoding is a popular statistical technique used in the credit rating industry [259]. It is used to automatically recode the values of continuous and categorical predictor variables into discrete bins, and to assign each bin a WoE value. The bins are determined such that they will produce the largest differences with respect to the WoE values. Additionally, monotonicity constraints can be specified to ensure that WoE values are strictly increasing or decreasing in feature values.

The formula for WoE encoding is derived from entropy theory and the information value. For bin_i, the WoE value can be computed as follows:

$$WoE_i = [\ln(\frac{\text{Relative Frequency of Goods}}{\text{Relative Frequency of Bads}})] * 100 \tag{G.1}$$

where the relative frequency of goods is defined as the ratio of the number of creditworthy individuals in bin_i to the total number of creditworthy individuals, and the relative frequency of bads is defined as the ratio of the number of non-creditworthy individuals in bin_i to the total number of non-creditworthy individuals.

Intuitively, the WoE value of a bin provides a measure of its predictive ability to separate creditworthy and non-creditworthy applicants. An important benefit of

WoE encoding is that it can be used to treat missing values and outliers without introducing sparsity. As WoE values are on the same scale, we can use them to compare the univariate effects of bins on the target variable within a feature or across all features. Its drawback, like most binning techniques, is that it results in a loss of information. As we will see in Section 9.5, models trained on WoE encoded data have a better performance than other methods.

G.1.5 Data Visualization

Before using the dataset to train machine learning models, we analyze its properties using a few exploratory data visualization techniques.

Figure G-2 visualizes the 23x23 correlation matrix of our dataset, identifying higher correlation values with lighter shades. We find three pairs of features with correlations greater than 0.8.

1. The total number of lines of credit (NumTotalLOC) and the number of lines of credit that are not currently delinquent (NumLOCNotDelq)

2. The number of lines of credit that have been 60+ days delinquent (NumLOC60PlusDaysDelq) and the number of lines of credit that have been 90+ days delinquent (NumLOC90PlusDaysDelq)

3. The number of request for new lines of credit in the past 6 months (NumLOCReqLast6M) and the number of request for new lines of credit in the past 6 months excluding the past week (NumLOCReqLast6MExPastWeek)

G.2 Model Implementation Details

The class distribution of the dataset in the five-fold cross-validation is presented in Table G.1. By visual inspection, it demonstrates that the dataset is reasonably balanced.

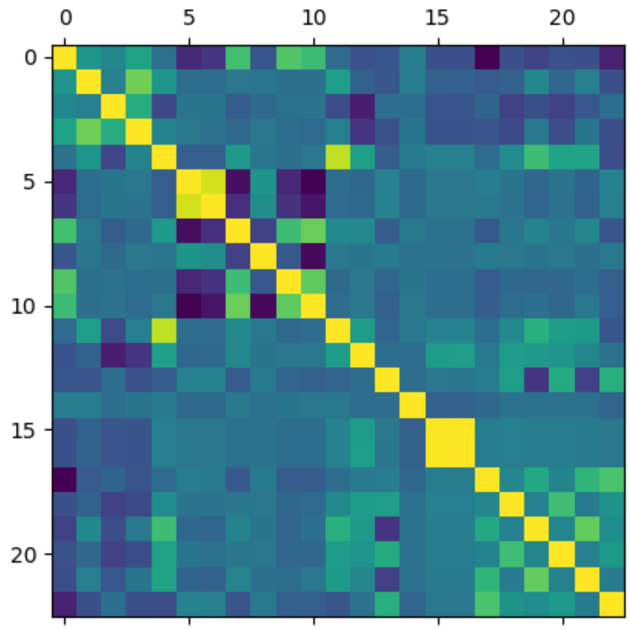


Figure G-2: 23 x 23 Correlation Matrix.
The lighter shades represents highly correlated feature pairs. As expected, the diagonal has the lightest shade because it represents the correlation of a feature with itself.

We give details of the implementation for the different ML models used in our evaluation.

For optimal trees, we used the Julia implementation available from the Interpretable AI website [1]. We performed a grid search over a depth from 1 to 10, and the number of features used for the deciding split at each node from the set 1,2,3,4,5,10,15,20,23. The optimal tree (black box) is the best-performing model of all the parameter combinations. Its parameters are depth=2 and the number of features=10. The optimal tree (interpretable) model corresponds to a model depth=2 and a number of features=1.

We used the scikit-learn implementation of the random forest classifier[2]. A grid search was performed on the number of estimators (trees) from 1 to 150. The best

[1] https://docs.interpretable.ai/v1.0/

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Fold Number	Count of $y_i = 1$ (Train)	Count of $y_i = 0$ (Train)	Count of $y_i = 1$ (Test)	Count of $y_i = 0$ (Test)
1	4118	3770	1010	963
2	4088	3801	1040	932
3	4094	3795	1034	938
4	4111	3778	1017	955
5	4101	3788	1027	945

Table G.1: Five-fold cross-validation data distribution. $y_i = 1$ corresponds to the data point (i.e., individual) who defaulted. The dataset is fairly balanced.

performing model consisted of 140 trees. Other parameters were set to the provided default values.

The neural network (WoE) model was implemented using the scikit-learn implementation of MLPClassifier. We performed a grid search over the different architectures and the L2-regularization constant. The best-performing model had three hidden layers consisting of 5 units each, and a L2-regularization penalty constant of 0.5. Other parameters were set to the default values.

The neural network (one-hot) model was implemented using tensorflow-keras. This was done to ensure the availability of gradients from the model for the counterfactual generating algorithm (DiCE). The best-performing model had one hidden layer, with the number of nodes=20. It was trained using the Adam optimizer with a learning rate of 0.01. The activation type used was the Rectified Linear Unit (ReLU), and a L1-regularization penalty of 0.001 was used.

We used publicly available implementation of SHAP[3]. We used publicly available implementation of LIME[4] and made appropriate modification to incorporate the feature correlation in the data sampling. We used publicly available implementation of DiCE[5] and made modification by specifying features that cannot be changed. We

[3] https://github.com/slundberg/shap
[4] https://github.com/marcotcr/lime
[5] https://github.com/interpretml/DiCE

implemented inductive logic programming using tensorflow-keras.

Appendix H

From ELIZA to ChatGPT: The Evolution of NLP and Financial Applications

H.1 Complete ChatGPT Interaction

One of the co-authors of this paper is ChatGPT, a chatbot developed by OpenAI and one of the current state-of-the art NLP tools. Here, we include the complete transcript of our conversation with ChatGPT that resulted in its contributions to this article.

Permission

AUTHORS: Can you write a research paper on evolution of NLP models and their use in finance?

CHATGPT: However, I can provide a brief overview of the evolution of NLP models and their use in finance:

Early NLP models: Early NLP models were based on rules-based systems and used hand-coded grammar rules to process language. These models were limited in their ability to handle complex language structures and were prone to error.

Statistical NLP models: With the advent of large amounts of data and computational power, statistical NLP models became popular. These models use statistical algorithms to analyze patterns in data and generate language outputs.

Deep Learning NLP models: With the development of deep learning algorithms, NLP models became more sophisticated. Deep learning models use neural networks to process and analyze language, allowing them to understand the meaning and context of language in a more nuanced way.

In finance, NLP models are used for a variety of purposes, including sentiment analysis, news summarization, and data analysis. NLP models can be used to analyze large amounts of financial news and information, identify trends and patterns, and provide insights into market conditions.

Overall, the evolution of NLP models has led to more advanced and sophisticated models that can better understand and process language, making them valuable tools for financial analysis and decision making.

ChatGPT was unable to write the full paper, but it gave a good overview of the topic. Following this, we started asking questions to ChatGPT step by step.

AUTHORS: I am writing a paper on evolution of NLP in finance. Can I use your help ?

CHATGPT: Yes, I'd be happy to help you with your paper on the evolution of NLP in finance. What specifically would you like to know or discuss?

Querying Time-line Information: Learning from Context

AUTHORS: My first question is, what were the earliest NLP based models in literature?

CHATGPT: The earliest NLP-based models in literature were rules-based systems, developed in the 1950s and 1960s. These systems utilized hand-crafted rules and grammars to analyze and understand natural language text. Later, statistical NLP models emerged in the 1980s and 1990s, which used machine learning algorithms to learn patterns in text data and make predictions about the meaning of text.

AUTHORS: When were the advanced statistical NLP models were developed ? Can you give me a timeline?

CHATGPT: A timeline of the development of advanced statistical NLP models is as follows:

Late 1990s to early 2000s: The first statistical NLP models, such as hidden Markov models (HMMs) and maximum entropy models, were developed and applied to various NLP tasks.

Mid 2000s to early 2010s: The widespread availability of large text corpora and advancements in machine learning algorithms led to the development of more sophisticated statistical NLP models, such as support vector machines (SVMs) and deep neural networks (DNNs).

Late 2010s to present: With the advent of big data and improvements in computational resources, deep learning techniques, such as recurrent neural networks (RNNs) and transformer models, have become the dominant NLP models and have been applied to a wide range of NLP tasks, including finance.

AUTHORS: I have heard about lexicon based models and topic models. When were they developed and they belong to which category of the NLP models

CHATGPT: "Lexicon-based" models and "topic models" are both categories of statistical NLP models.

Lexicon-based models: These models were developed in the early 2000s and rely on the use of predefined dictionaries or lexicons of words and phrases to analyze and understand natural language text. For example, a lexicon-based model might use a dictionary of financial terms to identify financial concepts in news articles.

Topic models: These models were developed in the late 1990s and early 2000s and use statistical techniques to identify the main topics discussed in a document or a collection of documents. For example, a topic model might identify market trends or company news as topics in financial news articles.

Both lexicon-based models and topic models are considered to be early forms of statistical NLP models and have been largely superseded by more advanced deep learning techniques, such as RNNs and transformers. However, they still play an important role in some NLP applications, particularly in the early stages of preprocessing and feature extraction.

AUTHORS: Can you give me a detailed timeline of the development of NLP models and their use in finance?

CHATGPT: A detailed timeline of the development of NLP models and their use in finance is as follows:

1950s to 1970s: Early NLP research focused on rules-based approaches, such as grammar-based parsers and dictionary-based information extraction systems.

Late 1980s to early 1990s: The first statistical NLP models, such as hidden Markov models (HMMs) and maximum entropy models (MEMs), were developed and applied to various NLP tasks.

Late 1990s to early 2000s: The development of lexicon-based models and topic models, which relied on predefined dictionaries or lexicons of words and phrases, represented an important step forward in NLP.

Mid 2000s to early 2010s: With the widespread availability of large text corpora and advancements in machine learning algorithms, more sophisticated statistical NLP models, such as support vector machines (SVMs) and deep neural networks (DNNs), were developed.

Late 2010s to present: With the advent of big data and improvements in computational resources, deep learning techniques, such as recurrent neural networks (RNNs) and transformer models, have become the dominant NLP models.

From the above snippet, we can observe that ChatGPT answer did not include the advanced statistical models like topic modeling or lexicon-based models initially, but on asking questions about the topic model and lexicon-based model, it included them in its answer to the next timeline question. This implies the chatbot is learning from context and correcting its mistakes.

AUTHORS: Can you give me some relevant research papers (citations) for the rules-based approaches like Grammar-based parsers, Dictionary-based information extraction, Text classification and Named entity recognition ?

CHATGPT: Sure, here are some relevant research papers for the rules-based approaches in NLP, such as grammar-based parsers, dictionary-based information extraction, text classification, and named entity recognition:

Grammar-based parsers:

Earley, J. (1970). An efficient context-free parsing algorithm. Communications of the ACM, 13(2), 94-102.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics-Volume 1.

Dictionary-based information extraction:

Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relationships from large plain-text collections. Proceedings of the sixth ACM SIGMOD Conference on Management of data.

Text classification:

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on Association for Computational Linguistics.

Named entity recognition:

R. Nadeau and S. Sekine. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

These are some of the classic and influential papers in the field of NLP and provide a good starting point for understanding the early rules-based approaches in NLP. However, it's worth noting that NLP research is a rapidly evolving field, so there are many other relevant and recent papers and advances that are not included in this list.

AUTHORS: Can you give me some relevant research papers (citations) for the rules-based approaches like Grammar-based parsers, Dictionary-based information extraction, Text classification and Named entity recognition from 1950-70s?

CHATGPT: Sure, here are some relevant research papers for the rules-based approaches in NLP from the 1950s to 1970s:

Grammar-based parsers:

Chomsky, N. (1957). Syntactic Structures. Mouton.

Ross, J. R. (1967). Constraints on variables in syntax. Doctoral dissertation, MIT.

Dictionary-based information extraction:

H. P. Luhn. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159-165.

Text classification:

G. H. W. Sperberg-McQueen. (1972). A natural language question-answering system. In Proceedings of the International Joint Conference on Artificial Intelligence, 397-406.

Named entity recognition:

S. S. Chhabra and J. E. Atwell. (1976). A study of the use of dictionary and grammar in the analysis of newswire texts. In Proceedings of the 10th International Conference on Computational Linguistics, 465-471.

These are some of the early and influential papers in the field of NLP from the 1950s to 1970s and provide a good starting point for understanding the rules-based approaches in NLP during that time. It's worth noting that the field of NLP has evolved significantly since then, and there are many other relevant and recent papers and advances that are not included in this list.

AUTHORS: Can you give me the above citations in the bibtex format ?

CHATGPT: Certainly, here are the citations in the BibTeX format for the rules-based approaches in NLP from the 1950s to 1970s: ....

From the above dialogue, it can be observed that ChatGPT understands fine details about our commands. On asking for references from the 1970s, it was able to change the references it offered and share ones from the 1970s instead. On asking a followup request to provide the citations in bibtex format, it was able to do so.

Following this interaction, we asked many questions related to different NLP models and their references.

Information/Content Generation

AUTHORS: Can you give me a detailed overview of the NLP models in 1980s-1990s:like Hidden Markov Models (HMMs), Maximum Entropy models, and decision trees?

CHATGPT: Sure, here's an overview of the NLP models in the 1980s-1990s:

Hidden Markov Models (HMMs): HMMs are a type of probabilistic model for modeling sequential data. They can be used for NLP tasks such as part-of-speech tagging, named entity recognition, and speech recognition. HMMs model sequences as a series of hidden states and observable outputs. They are trained using the Baum-Welch algorithm to estimate the model parameters that maximize the likelihood of the observed sequence.

Maximum Entropy Models: Maximum Entropy Models are probabilistic models that aim to maximize entropy (i.e., uncertainty) subject to constraints. They can be used for NLP tasks such as part-of-speech tagging, named entity recognition, and text classification. Maximum Entropy Models use a log-linear approach to model the probability of an output sequence given the input sequence. They are trained by maximizing the likelihood of the observed data subject to the constraints.

Decision Trees: Decision trees are a type of machine learning model that can be used for NLP tasks such as text classification and sentiment analysis. They model the decision process as a tree-like structure, where each node represents a decision based on the input features and each leaf node represents a prediction. Decision trees can be trained using algorithms such as ID3, C4.5, and CART. They are simple to interpret and can handle both continuous and categorical features.

Note: These are early models in NLP and there have been many advancements and improvements since then

AUTHORS: Can you give me citations and relevant papers on the use of Hidden Markov Models (HMMs) and Maximum Entropy models in NLP?

CHATGPT: Sure, here are some references and papers on the use of Hidden Markov Models (HMMs) and Maximum Entropy models in NLP:

Hidden Markov Models (HMMs):

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

Jiang, X. (2007). Part-of-speech tagging using hidden Markov models. In Natural language processing and knowledge engineering, 2007. NLP-KE'07. International Conference on (pp. 325-328). IEEE.

Maximum Entropy Models:

Berger, A. L. (1996). The maximum entropy approach to natural language processing. Computational Linguistics, 22(1), 39-71.

McCallum, A., & Li, W. (2003). Early results for named entity recognition with maximum entropy models. In Proceedings of the workshop on named entities: recognition, classification, and use in answer extraction (pp. 25-32).

Note: These are just a few references and there are many other papers and studies that cover the use of HMMs and Maximum Entropy models in NLP.

AUTHORS: What were the limitation of Hidden Markov Models (HMMs) and Maximum Entropy Models in NLP ?

CHATGPT: Here are some of the limitations of Hidden Markov Models (HMMs) and Maximum Entropy Models in NLP:

Hidden Markov Models (HMMs): HMMs have difficulty in handling long-range dependencies, as they only consider local context. HMMs also have difficulties in dealing with ambiguity and uncertainty, as they assume that the observations are generated from a single hidden state.

Maximum Entropy Models: Maximum Entropy Models require a large amount of training data in order to produce accurate results. They also have difficulty in modeling complex relationships between variables, as they assume independence between variables. Maximum Entropy Models also struggle with computational efficiency, as they require large amounts of computing resources to estimate parameters.

Note: These limitations were more pronounced in the early days of NLP and have been addressed to some extent with the development of more advanced NLP models. However, these limitations still persist to some extent in modern NLP models.

AUTHORS: Can you give me some influential papers on on lexicon based model and topic modeling in NLP?

CHATGPT: Sure, here are some influential papers on lexicon-based models and topic modeling in NLP:

Lexicon-based Models:

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing.

Huang, J., Zhang, S., & Zhao, Y. (2012). Improving sentiment classification via sentiment concept clustering and transfer learning. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.

Topic Modeling:

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Asuncion, A., Welling, M., Smola, A. J., & Teh, Y. W. (2009). On smoothing and inference for topic models. Proceedings of the 26th Annual International Conference on Machine Learning.

Note: These are just a few examples of influential papers on lexicon-based models and topic modeling in NLP and there are many other papers that have made significant contributions to these areas.

AUTHORS: Can you give me overview of the word embeddings based methods in NLP and their use in finance ?

CHATGPT: Word embeddings are a type of NLP technique that represent words as continuous-valued vectors in a high-dimensional space. These word vectors capture semantic and syntactic information about words in a way that can be used for various NLP tasks, including sentiment analysis, text classification, and language translation.

In finance, word embeddings have been used for tasks such as:

Sentiment Analysis: Word embeddings have been used to analyze the sentiment of financial news articles, investor reports, and social media posts, in order to gauge market sentiment and predict stock prices.

Text Classification: Word embeddings have been used to classify financial documents into categories such as earnings reports, balance sheets, and press releases.

Risk Assessment: Word embeddings have been used to analyze the risk associated with financial products, by identifying potential red flags in regulatory reports and other financial documents.

Topic Modeling: Word embeddings have been used in topic modeling of financial documents, in order to identify important themes and topics in financial data.

Overall, word embeddings have proven to be a powerful tool for NLP in finance, as they enable the use of high-level abstractions to analyze large amounts of unstructured financial data, and provide insights into market sentiment and risk.

AUTHORS: Can you give me citations for the above methods in bibtex format?

CHATGPT: I'm sorry, I cannot provide citations for all the word embedding methods listed in my previous response. However, here are some papers that cover popular word embedding methods: ...

AUTHORS: What are the disadvantages of embedding based methods in NLP?

CHATGPT: Some disadvantages of word embedding based methods in NLP include:

Data sensitivity: word embeddings are highly dependent on the training corpus, and can be sensitive to the type of data and domain they are trained on.

Out-of-vocabulary (OOV) words: word embeddings do not provide good representations for words that are not present in the training corpus, and often perform poorly on OOV words.

Context sensitivity: word embeddings are context sensitive, and the representation for a word can change depending on the context it is used in. This can lead to issues with polysemy (multiple meanings for a single word) and synonymy (multiple words with the same meaning).

Scalability: computing word embeddings for large corpora can be computationally intensive, and can require significant computational resources.

Semantic compositionality: word embeddings do not directly capture the compositional nature of language, and it can be challenging to use them to effectively capture the meaning of phrases and sentences.

Overfitting: word embeddings can overfit to the training corpus, especially when trained on small data. This can result in the representations capturing irrelevant information or biases present in the training data.

AUTHORS: What are the new deep learning based NLP methods?

CHATGPT: New deep learning based NLP methods include:

Transformer models: these models have revolutionized NLP by introducing the self-attention mechanism, which allows for parallel processing of input sequences. Examples include BERT (Bidirectional Encoder Representations from Transformers), GPT-2/3 (Generative Pretrained Transformer 2/3), and RoBERTa (Robustly Optimized BERT Approach).

Neural Machine Translation (NMT): these models use deep neural networks to perform machine translation, and have shown significant improvement over traditional statistical machine translation methods.

Seq2Seq models: these models are designed for sequence-to-sequence tasks such as machine translation, text summarization, and text generation. They consist of an encoder-decoder architecture, where the encoder maps an input sequence to a fixed-length vector, and the decoder generates an output sequence from the vector.

Convolutional Neural Networks (CNNs): these models are designed for text classification and sentiment analysis tasks, and have shown good performance on a wide range of NLP tasks.

Recurrent Neural Networks (RNNs): these models are designed for sequential data, and have been widely used in NLP for tasks such as language modeling, sentiment analysis, and text generation.

Generative Adversarial Networks (GANs): these models are used to generate synthetic text data, and have shown good performance in tasks such as text generation, data augmentation, and text style transfer.

These deep learning based NLP methods have shown remarkable progress and have been widely adopted in NLP research and industry applications.

AUTHORS: What are the disadvantages of deep learning based NLP models?

CHATGPT: Some disadvantages of deep learning based NLP models include:

Computational cost: deep learning based NLP models are computationally intensive, and require large amounts of data, memory, and computational resources to train.

Data bias: deep learning models can perpetuate the biases present in the training data, leading to unfair or undesirable outcomes in real-world applications.

Overfitting: deep learning models can easily overfit to the training data, especially when trained on small or noisy data. This can result in poor generalization performance and low accuracy on unseen data.

Explanation difficulty: deep learning models are often considered to be black boxes, with little understanding of how they make predictions. This can make it challenging to diagnose and fix issues with the models, and to build trust in the results they produce.

Transfer learning difficulty: deep learning models are often trained on large, diverse corpora, making it difficult to fine-tune them for specific domains or tasks without access to large amounts of annotated data.

Adversarial examples: deep learning models are vulnerable to adversarial examples, where small changes to the input data can lead to incorrect predictions. This can be a major security concern in NLP applications where the input data can be easily manipulated.

Overall, deep learning based NLP models have shown great progress, but also have some significant limitations that need to be considered when designing and deploying NLP systems.

AUTHORS: Can you give me citations for the influential papers related to the deep learning based NLP models ?

CHATGPT: Here are some influential papers in the field of deep learning based NLP models, with bibtex citations:

BERT: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin et al., 2018. ....

From the above conversation, we obtained a significant amount of information content. However, there were a few times when its response was not very detailed or relevant to the question. For such replies, we had to select the relevant information. We also observed that from its history ChatGPT understood the importance of giving a citation in bibtex format, hence in reply to the last question, it automatically returned the citation in bibtex format.

Questions About ChatGPT

AUTHORS: What is ChatGPT?

CHATGPT: GPT-3 (Generative Pretrained Transformer 3) is a language model developed by OpenAI, a research organization dedicated to creating and promoting artificial intelligence in a responsible way. It is the third iteration in the GPT series of models, and the largest one to date, with over 175 billion parameters.

GPT-3 is trained on a massive dataset of internet text and uses a deep neural network architecture based on the Transformer architecture. The model is capable of performing various natural language processing tasks, such as text generation, question answering, machine translation, and sentiment analysis, among others.

ChatGPT is a specific implementation of GPT-3 designed for conversational AI applications. It is fine-tuned to respond to user inputs in a conversational manner, and can be integrated into chatbots and other conversational applications to provide human-like responses.

AUTHORS: Can I get more details about the ChatGPT ?

CHATGPT: Sure! Here's some more information about ChatGPT:

Architecture: ChatGPT is based on the transformer architecture, which is a deep neural network designed for processing sequential data, such as natural language. The transformer architecture is made up of multiple layers, including an attention mechanism, which allows the model to focus on different parts of the input at different times and dynamically adjust its internal representations.

Pretraining: Like other GPT models, ChatGPT is pretrained on a large corpus of text data. During pretraining, the model learns to predict the next word in a sentence given its preceding context. This enables the model to build a rich representation of the language and learn patterns and relationships between words and phrases.

Fine-tuning: Once the model is pretrained, it can be fine-tuned for specific tasks and domains, such as conversational AI. During fine-tuning, the model is trained on a smaller dataset that is relevant to the target task. This allows the model to learn task-specific information and improve its performance on the target task.

Applications: ChatGPT is used in a variety of conversational AI applications, including chatbots, virtual assistants, and customer service systems. The model can be used to generate human-like responses to user inputs, answer questions, provide recommendations, and perform other conversational tasks.

Limitations: Although ChatGPT is a highly advanced and powerful language model, it is not perfect. Like all machine learning models, it can make mistakes and produce responses that are irrelevant, incorrect, or offensive. Additionally, its large size and computational requirements make it difficult to deploy in real-world applications, especially in resource-constrained environments.

AUTHORS: What is the citation for ChatGPT?

CHATGPT: ChatGPT is a language model developed by OpenAI, and as such, it does not have a specific publication or citation. OpenAI is a research organization that develops and releases a number of advanced AI models, including GPT-3, which is the basis for ChatGPT. You can learn more about OpenAI and its research on their website (https://openai.com/).

AUTHORS: How much energy ChatGPT uses for generation of 100 words? The energy usage for generating 100 words of text with ChatGPT would depend on several factors, including the hardware and software configurations used, the size of the model, and the computational resources available.

As a large language model, ChatGPT requires significant computational resources, including a large amount of memory and processing power. The exact energy usage would therefore vary depending on the specific hardware and software environment in which the model is running.

It is worth noting that the energy usage of large language models like ChatGPT has been a topic of concern in the AI community due to the significant carbon footprint associated with training and using such models. Research is ongoing to find ways to reduce the energy usage of these models and make them more environmentally sustainable.

In the above conversation, we asked questions that related ChatGPT to itself, and it was able to return detailed information. Hence, we conclude that ChatGPT is self-aware. However, it was not able to reply with technical details about the energy used by ChatGPT.

Information/Content Generation

AUTHORS: Can you give me examples of use of NLP for risk management in finance?

CHATGPT: Yes, here are some examples of the use of NLP in risk management in finance:

Sentiment analysis: NLP techniques can be used to analyze large volumes of unstructured text data, such as financial news articles, social media posts, and earnings call transcripts, to gain insights into market sentiment and assess the risk of specific financial instruments.

News and event analysis: NLP models can be trained to identify and classify relevant financial news and events and predict their impact on the market and specific financial instruments.

Regulatory compliance: NLP can be used to automatically scan financial documents and identify potential compliance violations, such as insider trading or money laundering, to mitigate financial risk.

Credit scoring: NLP can be used to analyze financial statements and other financial data to predict a company's credit risk and support decision-making in lending and credit management.

Fraud detection: NLP can be used to analyze transaction data and detect suspicious patterns that may indicate fraud, helping financial institutions minimize their risk exposure.

These are just a few examples of the many ways NLP can be used for risk management in finance.

AUTHORS: Can you give me examples of use of NLP for impact investing in finance?

CHATGPT: Use of NLP to analyze and categorize sustainability-related information contained in companies' annual reports and sustainability disclosures.

NLP-based sentiment analysis of news articles and social media posts to track and quantify the social and environmental impact of companies.

NLP-based event extraction to monitor and quantify ESG-related events and their impact on companies and industries.

NLP-based entity extraction to identify companies and projects with a positive impact on society and the environment.

NLP-based document classification to automatically categorize sustainability reports and disclosures according to specific impact themes, such as climate change, human rights, and biodiversity.

NLP-based natural language generation to produce impact reports and summaries that highlight a company's positive and negative impact on society and the environment.

AUTHORS: What is risk management in finance ?

CHATGPT: Risk management in finance refers to the process of identifying, assessing, and controlling risks in financial operations, transactions, and investments. It involves identifying potential sources of financial loss and taking steps to minimize or mitigate their impact on the financial stability and performance of an organization. Key components of financial risk management include market risk, credit risk, operational risk, liquidity risk, and reputational risk. The goal of financial risk management is to minimize the impact of adverse events on an organization's financial performance and stability.

AUTHORS: Can you give me more examples of influential papers that use of NLP in risk management in finance ?

CHATGPT: I apologize, but I do not have specific references to influential papers that use NLP in risk management in finance at this time.

AUTHORS: What is impact investing ?

CHATGPT: Impact investing is a form of investment that aims to generate positive social or environmental impact alongside financial returns. Impact investments are made into companies, organizations, and funds with the intention to generate measurable social and environmental impact alongside financial returns. This investment approach seeks to create social or environmental benefits alongside financial returns, as opposed to philanthropy which aims primarily to create social or environmental benefits.

AUTHORS: Can you give me detailed summary of use of NLP for asset management in finance?

CHATGPT: Natural Language Processing (NLP) has been used in asset management for various tasks such as sentiment analysis, news analysis, and earnings call analysis.

Sentiment analysis involves analyzing the tone and emotion of financial news articles and social media posts to predict stock market trends. NLP models are used to process and analyze vast amounts of text data and extract meaningful information to make investment decisions.

News analysis involves using NLP to extract relevant information from news articles, such as company news, earnings releases, and analyst reports. This information can be used to make investment decisions and improve the portfolio management process.

Earnings call analysis involves using NLP to process and analyze earnings call transcripts. The goal is to extract key financial metrics and sentiment indicators to better understand a company's financial performance and make investment decisions.

However, despite its potential, the use of NLP in asset management also faces some limitations such as the difficulty in accurately interpreting text data, the high computational cost of training NLP models, and the need for large amounts of labeled data to train the models.

AUTHORS: what is the future of NLP ?

CHATGPT: The future of NLP looks promising with continued advancements in deep learning techniques and increasing availability of data. Advancements in NLP are likely to improve areas such as natural language generation, dialogue systems, and multi-language processing. Additionally, NLP applications are expected to expand beyond the traditional domains of information retrieval and text classification, to more complex real-world problems, such as recommendation systems, personalization, and data privacy. The integration of NLP with other areas, such as reinforcement learning, computer vision, and robotics, is also likely to drive further innovation. However, the continued development of NLP technology will also bring up new ethical and societal issues, such as data privacy, bias, and fairness.

Creativity

AUTHOR: What are the suitable titles for the paper on evolution of NLP in finance ?

CHATGPT: "From Text Mining to Intelligent Finance: A Journey Through NLP"

"The Rise of NLP in Financial Services: Opportunities and Challenges"

"NLP in Finance: Past, Present, and Future Trends"

"Leveraging Natural Language Processing for Improved Financial Outcomes"

"Transforming Finance with NLP: A Review of the Literature"

"NLP and Financial Analytics: A Study of Emerging Trends and Techniques"

"The Power of NLP in Financial Decision Making: A Review"

"NLP and Financial Risk Management: Progress and Prospects"

"Enhancing Financial Performance with NLP: A Survey of Applications"

"Natural Language Processing and Finance: A Comprehensive Review".

AUTHOR: Can you give me famous quotes for the use of NLP in finance?

CHATGPT: Here are a few famous quotes related to the use of NLP in finance:

"Data is the new oil and NLP is the refinery." - Anonymous

"The future of finance lies in the intersection of machine learning and NLP." - William Falcon, CEO of Fast.ai

"NLP is revolutionizing finance by providing insights from unstructured data." - Julie Delanoy, CEO of Cognitivescale

"The power of NLP lies in its ability to turn data into actionable insights for finance professionals." - John Giannandrea, CEO of GPT-3

"The use of NLP in finance is providing a more efficient and accurate way to analyze large amounts of data." - Richard Brown, CTO of AI at Accenture

“NLP is transforming finance by providing a more human-like understanding of financial data." - Oren Etzioni, CEO of the Allen Institute for Artificial Intelligence

AUTHORS: Can you give me some more famous quotes for the use of NLP in finance?

CHATGPT: I'm sorry, I don't have a database of famous quotes specifically related to the use of NLP in finance. Quotes are often subjective and personal, so it may be best to create your own based on your own experiences and opinions.

AUTHORS: Whats the future of NLP in finance ?

CHATGPT: The future of NLP in finance looks bright as NLP has already demonstrated its ability to perform tasks such as sentiment analysis, risk management, and impact investing, among others. With advancements in NLP technology and the increasing volume of financial data available, NLP is poised to play an increasingly important role in finance, with potential applications ranging from financial analysis and investment decision-making to fraud detection and regulatory compliance. However, it is important to consider the limitations of NLP, such as data scarcity problems, ethical and legal implications, and limitations of NLP models themselves. Overall, the future of NLP in finance will likely be shaped by a combination of technological progress and practical considerations, as the financial industry continues to adopt NLP and reap the benefits of its powerful capabilities

From the above dialogue with ChatGPT, we can observe that the chatbot is creative, as it is able to provide us with suitable titles for this paper, along with quotes related to NLP in finance.

Logical Reasoning

In this section, we discuss an example from ChatGPT where it failed to answer a simple logical reasoning question without a hint.

AUTHORS: I am A's grandson. A have two sons, B and C. B loves dogs. I love cats. B doesn't have children. My brother lives in San Francisco. Who is my father?

CHATGPT: If you are A's grandson, it means that A is your grandfather. It is not clear from the information provided who your father is, as the information only states that A has two sons, B and C, and does not mention which son is your father.

AUTHORS: There is enough information is you consider the fact that B does not have children.

CHATGPT: I apologize for my previous response. If B is A's son and he does not have any children, then C must be your father since A has two sons, and you are B's son and you mentioned that B does not have any children. Therefore, your father must be C.

Bibliography

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] Sumit Agarwal, Gene Amromin, Itzhak Ben-David, Souphala Chomsisengphet, and Douglas D Evanoff. The role of securitization in mortgage renegotiation. *Journal of Financial Economics*, 102(3):559–578, 2011.

[3] Sumit Agarwal, Yan Chang, and Abdullah Yavas. Adverse selection in mortgage securitization. *Journal of Financial Economics*, 105(3):640–660, 2012.

[4] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94, 2000.

[5] Shreyash Agrawal, Pablo D Azar, Andrew W Lo, and Taranjit Singh. Momentum, mean-reversion, and social media: Evidence from stocktwits and twitter. *The Journal of Portfolio Management*, 44(7):85–95, 2018.

[6] Daniel Felix Ahelegbey, Paolo Giudici, and Branka Hadji-Misheva. Latent factor models for credit scoring in p2p systems. *Physica A: Statistical Mechanics and its Applications*, 522:112–121, 2019.

[7] Fares Al-Shargie, Masashi Kiguchi, Nasreen Badruddin, Sarat C Dass, Ahmad Fadzil Mohammad Hani, and Tong Boon Tang. Mental stress assessment using simultaneous measurement of eeg and fnirs. *Biomedical optics express*, 7(10):3882–3898, 2016.

[8] Stefania Albanesi and Domonkos F Vamossy. Predicting consumer default: A deep learning approach. Technical report, National Bureau of Economic Research, 2019.

[9] Ana Albuquerque. Peer firms in relative performance evaluation. *Journal of Accounting and Economics*, 48(1):69–89, 2009.

[10] Rui Albuquerque, Yrjö Koskinen, and Chendi Zhang. Corporate social responsibility and firm risk: Theory and empirical evidence. *Management Science*, 65(10):4451–4469, 2019.

[11] Basemah Alshemali and Jugal Kalita. Improving the reliability of deep neural networks in nlp: A review. *Knowledge-Based Systems*, 191:105210, 2020.

[12] Werner Antweiler and Murray Z Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294, 2004.

[13] Coren L Apicella, Anna Dreber, Benjamin Campbell, Peter B Gray, Moshe Hoffman, and Anthony C Little. Testosterone and financial risk preferences. *Evolution and human behavior*, 29(6):384–390, 2008.

[14] Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020.

[15] Miller Janny Ariza-Garzón, Javier Arroyo, Antonio Caparrini, and Maria-Jesus Segovia-Vargas. Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access*, 8:64873–64890, 2020.

[16] Arthur Asuncion, Max Welling, Alexander J Smola, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[17] Pablo D Azar and Andrew W Lo. The wisdom of twitter crowds: Predicting stock market reactions to fomc meetings via twitter feeds. *The Journal of Portfolio Management*, 42(5):123–134, 2016.

[18] J.K. Baker. Trainable grammars for speech recognition. In *The Journal of the Acoustical Society of America*, volume 66, pages 1333–1344, 1979.

[19] Scott R Baker, Nicholas Bloom, Steven J Davis, Kyle J Kost, Marco C Sammon, and Tasaneeya Viratyosin. The unprecedented stock market impact of covid-19. Technical report, national Bureau of economic research, 2020.

[20] Anjana Bali and Amteshwar Singh Jaggi. Clinical experimental stress studies: methods and assessment. *Reviews in the Neurosciences*, 26(5):555–579, 2015.

[21] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[22] M Bayraktar, I Mashtaler, N Meng, and S Radchenko. Barra us total market equity trading model empirical notes. *MSCI Model Insight*, 2015.

[23] Chinmaya Behera and Badri Narayan Rath. The covid-19 pandemic and indian pharmaceutical companies: An event study analysis. *Bulletin of Monetary Economics and Banking, Special*, (2021):1–14, 2021.

[24] Florian Berg, Kornelia Fabisik, and Zacharias Sautner. Is history repeating itself? the (un) predictable past of esg ratings. *The (Un) Predictable Past of ESG Ratings (August 24, 2021). European Corporate Governance Institute–Finance Working Paper*, 708, 2020.

[25] Florian Berg, Florian Heeb, and Julian F Kölbel. The economic impact of esg rating changes. *Available at SSRN 4088545*, 2022.

[26] Florian Berg, Julian F Kölbel, Anna Pavlova, and Roberto Rigobon. ESG confusion and stock returns: Tackling the problem of noise. *Available at SSRN 3941514*, 2021.

[27] Florian Berg, Julian F Kölbel, and Roberto Rigobon. Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6):1315–1344, 05 2022.

[28] Adam L. Berger. The maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[29] G.R. Bergersen. Maximum entropy models in natural language processing. In *Proceedings of the International Conference on Computational Linguistics*, pages 267–274, 1994.

[30] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

[31] PK Bhattacharya. Convergence of sample paths of normalized sums of induced order statistics. *The Annals of Statistics*, 2(5):1034–1039, 1974.

[32] Sanjeev Bhojraj, Charles MC Lee, and Derek K Oler. What's my line? a comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, 41(5):745–774, 2003.

[33] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[34] Patrick Bolton and Marcin T Kacperczyk. Carbon premium around the world. CEPR Discussion Paper No. DP14567, Available at SSRN: https://ssrn.com/abstract=3594188, 2020.

[35] George Bonne, Andrew W. Lo, Abilash Prabhakaran, Kien Wei Siah, Manish Singh, Xinxin Wang, Peter Zangari, and Howard Zhang. An artificial intelligence-based industry peer grouping system. *The Journal of Financial Data Science*, 4(2):9–36, 2022.

[36] Christa HS Bouwman, Kathleen Fuller, and Amrita S Nain. Market valuation and acquisition quality: Empirical evidence. *The Review of Financial Studies*, 22(2):633–679, 2009.

[37] Ronald J Brachman, Tom Khabaza, Willi Kloesgen, Gregory Piatetsky-Shapiro, and Evangelos Simoudis. Mining business databases. *Communications of the ACM*, 39(11):42–48, 1996.

[38] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. Machine learning explainability in finance: an application to default risk analysis. Bank of England working papers 816, Bank of England, August 2019.

[39] Joseph L Breeden. Survey of machine learning in credit risk. *Available at SSRN 3616342*, 2020.

[40] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[41] Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[42] Stephen J Brown and Jerold B Warner. Using daily stock returns: The case of event studies. *Journal of financial economics*, 14(1):3–31, 1985.

[43] Matthias MM Buehlmaier and Toni M Whited. Are financial constraints priced? evidence from textual analysis. *The Review of Financial Studies*, 31(7):2693–2728, 2018.

[44] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. Explainable machine learning in credit risk management. *Computational Economics*, pages 1–14, 2020.

[45] Florentin Butaru, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W Lo, and Akhtar Siddique. Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72:218–239, 2016.

[46] Colin F Camerer, George Loewenstein, and Matthew Rabin. *Advances in behavioral economics*. Princeton university press, 2004.

[47] John Y Campbell, Andrew W Lo, A Craig MacKinlay, and Robert F Whitelaw. The econometrics of financial markets. *Macroeconomic Dynamics*, 2(4):559–562, 1998.

[48] Tim S Campbell and J Kimball Dietrich. The determinants of default on insured conventional residential mortgage loans. *The Journal of Finance*, 38(5):1569–1581, 1983.

[49] Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, 19(8):1849, 2019.

[50] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[51] Louis KC Chan, Josef Lakonishok, and Bhaskaran Swaminathan. Industry classifications and return comovement. *Financial Analysts Journal*, 63(6):56–70, 2007.

[52] Wesley S Chan. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260, 2003.

[53] Paul K Chaney and Timothy M Devinney. New product innovations and stock price performance. *Journal of Business Finance & Accounting*, 19(5):677–695, 1992.

[54] John M Chapman. Factors affecting credit risk in personal lending. In *Commercial Banks and Consumer Instalment Credit*, pages 109–139. NBER, 1940.

[55] Anusha Chari, Paige Ouimet, and Linda L Tesar. Cross border mergers and acquisitions in emerging markets: The stock market valuation of corporate control. *Available at SSRN 559923*, 2004.

[56] Caroline J. Charpentier, Jan-Emmanuel De Neve, Xinyi Li, Jonathan P. Roiser, and Tali Sharot. Models of affective decision making: How do feelings predict choice? *Psychological science*, 27(6):763—-775, 2016.

[57] Shomesh Chaudhuri, Katherine Cheng, Andrew W Lo, Shirley Pepke, Sergio Rinaudo, Lynda Roman, and Ryan Spencer. A portfolio approach to accelerate therapeutic innovation in ovarian cancer. *Available at SSRN 3286330*, 2018.

[58] Sudheer Chava. Environmental externalities and cost of capital. *Management science*, 60(9):2223–2247, 2014.

[59] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. *Decision Support Systems*, 152:113647, 2022.

[60] Shunqin Chen, Zhengfeng Guo, and Xinlei Zhao. Predicting mortgage early delinquency with machine learning methods. *European Journal of Operational Research*, 290(1):358–372, 2021.

[61] Yu Chen, Rhaad M Rabbani, Aparna Gupta, and Mohammed J Zaki. Comparative text analytics via topic modeling in banking. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE, 2017.

[62] Youngjun Cho, Simon J Julier, and Nadia Bianchi-Berthouze. Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. *JMIR mental health*, 6(4):e10140, 2019.

[63] N. Chomsky. *Syntactic Structures*. Mouton, 1957.

[64] Brian Christian. The samantha test. *New Yorker*, 2013.

[65] George P Chrousos. Stress and disorders of the stress system. *Nature reviews endocrinology*, 5(7):374–381, 2009.

[66] John Coates and Mark Gurnell. Combining field work and laboratory work in the study of financial risk-taking. *Hormones and behavior*, 92:13–19, 2017.

[67] John M Coates and Joe Herbert. Endogenous steroids and financial risk taking on a london trading floor. *Proceedings of the national academy of sciences*, 105(16):6167–6172, 2008.

[68] Theodor Cojoianu, Andreas GF Hoepner, Georgiana Ifrim, and Yanan Lin. Greenwatch-shing: Using ai to detect greenwashing. *AccountancyPlus-CPA Ireland*, 2020.

[69] Charles J Corrado. Event studies: A methodology review. *Accounting & Finance*, 51(1):207–234, 2011.

[70] Raritan Costin, Cristian Rotariu, and Alexandru Pasarica. Mental stress detection using heart rate variability and morphologic variability of eeg signals. In *2012 International Conference and Exposition on Electrical and Power Engineering*, pages 591–596. IEEE, 2012.

[71] Matteo Cristofaro, Pier Luigi Giardino, Andrea P. Malizia, and Antonio Mastrogiorgio. Affect and cognition in managerial decision making: A systematic literature review of neuroscience evidence. *Frontiers in Psychology*, 13, 2022.

[72] K Croxson, P Bracke, and C Jung. Explaining why the computer says 'no'. *FCA*, 5:31, 2019.

[73] Carlos Cueva, R Edward Roberts, Tom Spencer, Nisha Rani, Michelle Tempest, Philippe N Tobler, Joe Herbert, and Aldo Rustichini. Cortisol and testosterone increase financial risk taking and may destabilize markets. *Scientific reports*, 5(1):1–16, 2015.

[74] Donald F Cunningham and Patric H Hendershott. Pricing fha mortgage default insurance. Technical report, National Bureau of Economic Research, 1984.

[75] Andrew M Dai, Christopher Olah, and Quoc V Le. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*, 2015.

[76] Margaret Dalziel, Xiangyang Yang, Simon Breslav, Azam Khan, and Jianxi Luo. Can we design an industry classification system that reflects industry architecture? *Journal of Enterprise Transformation*, 8(1-2):22–46, 2018.

[77] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, 2020.

[78] Sanjiv R Das and Mike Y Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management science*, 53(9):1375–1388, 2007.

[79] Sonya Das, Raphaël Rousseau, Peter C Adamson, and Andrew W Lo. New business models to accelerate innovation in pediatric oncology therapeutics: a review. *JAMA oncology*, 4(9):1274–1280, 2018.

[80] Sanjeeb Dash, Oktay Günlük, and Dennis Wei. Boolean decision rules via column generation. *arXiv preprint arXiv:1805.09901*, 2018.

[81] Randall Davis, Andrew W Lo, Sudhanshu Mishra, Arash Nourian, Manish Singh, Nicholas Wu, and Ruixun Zhang. Explainable machine learning models of consumer credit risk. *Available at SSRN*, 2022.

[82] Eric De Bodt, B Espen Eckbo, and Richard Roll. Corporate rivalry and return comovement. *Tuck School of Business Working Paper*, (3218544), 2020.

[83] Benedetto De Martino, John P O'Doherty, Debajyoti Ray, Peter Bossaerts, and Colin Camerer. In the mind of the market: Theory of mind biases value computation during financial bubbles. *Neuron*, 79(6):1222–1231, 2013.

[84] Yongheng Deng, John M Quigley, and Robert Van Order. Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2):275–307, 2000.

[85] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[86] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

[87] Chukwuma Dim, Kevin Koerner, Marcin Wolski, and Sanne Zwart. Hot off the press: News-implied sovereign default risk. Technical report, EIB Working Papers, 2022.

[88] Roger Dobson. Biotechnology company's shares dive 67% after drug failure, 2000.

[89] James Clay Dolley. Characteristics and procedure of common stock split-ups. *Harvard business review*, 11(3):316–326, 1933.

[90] Julian Donovan. The impact of drug and firm attributes on stock price movements around fda drug approval decisions. Technical report, Western University, 2018.

[91] Gregor Dorfleitner, Gerhard Halbritter, and Mai Nguyen. Measuring the level and risk of corporate responsibility–an empirical comparison of different esg rating approaches. *Journal of Asset Management*, 16(7):450–466, 2015.

[92] Casey Dougal, Joseph Engelberg, Diego Garcia, and Christopher A Parsons. Journalists and the stock market. *The Review of Financial Studies*, 25(3):639–679, 2012.

[93] Jing Duan. Financial system modeling using deep neural networks (dnns) for effective risk assessment and prediction. *Journal of the Franklin Institute*, 356(8):4716–4731, 2019.

[94] Elena Dumitrescu, Sullivan Hué, Christophe Hurlin, and Sessi Tokpavi. Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds. Working Papers hal-02507499, HAL, January 2021.

[95] J. Earley. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102, 1970.

[96] Alex Edmans. Does the stock market fully value intangibles? employee satisfaction and equity prices. *Journal of Financial economics*, 101(3):621–640, 2011.

[97] Alex Edmans, Diego García, and Øyvind Norli. Sports sentiment and stock returns. *The Journal of Finance*, 62(4):1967–1998, 2007.

[98] Sadok El Ghoul, Omrane Guedhami, Chuck CY Kwok, and Dev R Mishra. Does corporate social responsibility affect the cost of capital? *Journal of banking & finance*, 35(9):2388–2406, 2011.

[99] Ronel Elul, Nicholas S Souleles, Souphala Chomsisengphet, Dennis Glennon, and Robert Hunt. What" triggers" mortgage default? *American Economic Review*, 100(2):490–94, 2010.

[100] Scott R Evans. Fundamentals of clinical trial design. *Journal of experimental stroke & translational medicine*, 3(1):19, 2010.

[101] David E Fagnan, Austin A Gromatzky, Roger M Stein, Jose-Maria Fernandez, and Andrew W Lo. Financing drug discovery for orphan diseases. *Drug discovery today*, 19(5):533–538, 2014.

[102] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[103] Eugene F Fama, Lawrence Fisher, Michael Jensen, and Richard Roll. The adjustment of stock prices to new information. *International economic review*, 10(1), 1969.

[104] Eugene F. Fama and Kenneth R. French. Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193, 1997.

[105] Eugene F Fama and Kenneth R French. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22, 2015.

[106] Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, 2020.

[107] James R Ferguson. Biomedical research and insider trading, 1997.

[108] Javier Martínez Fernández, Ralf Seepold, Juan Carlos Augusto, and Natividad Martínez Madrid. Sensors in trading process: A stress—aware trader. In *2010 8th Workshop on Intelligent Solutions in Embedded Systems*, pages 17–22. IEEE, 2010.

[109] Jose-Maria Fernandez, Roger M Stein, and Andrew W Lo. Commercializing biomedical research through securitization techniques. *Nature biotechnology*, 30(10):964–975, 2012.

[110] Michael Firth. The impact of earnings announcements on the share price behaviour of similar type firms. *The Economic Journal*, 86(342):296–306, 1976.

[111] Dov Fischer. Investor underreaction to earnings surprises and overreaction to product news in the drug industry. *The Journal of Business and Economic Studies*, 18(2):82, 2012.

[112] Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214, 2016.

[113] Christopher Foote, Kristopher Gerardi, Lorenz Goette, and Paul Willen. Reducing foreclosures: No easy answers. *NBER Macroeconomics Annual*, 24(1):89–138, 2010.

[114] Katherine Beal Frazier, Robert W Ingram, and B Mack Tennyson. A methodology for the analysis of narrative accounting disclosures. *Journal of Accounting Research*, pages 318–331, 1984.

[115] Cary Frydman, Nicholas Barberis, Colin Camerer, Peter Bossaerts, and Antonio Rangel. Using neural data to test a theory of investor behavior: An application to realization utility. *The Journal of finance*, 69(2):907–946, 2014.

[116] Cary Frydman and Colin F Camerer. The psychology and neuroscience of financial decision making. *Trends in cognitive sciences*, 20(9):661–675, 2016.

[117] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022.

[118] Martin Garriga, Stefano Dalla Palma, Maxmiliano Arias, Alan De Renzis, Remo Pareschi, and Damian Andrew Tamburri. Blockchain and cryptocurrencies: A classification and comparison of architecture drivers. *Concurrency and Computation: Practice and Experience*, 33(8):e5992, 2021.

[119] Sebastien Gay and Ezra Karger. Patent citations and stock performance: Constructing a dynamic industry classification. *Available at SSRN 2496414*, 2014.

[120] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 2019.

[121] Rajna Gibson Brandon, Philipp Krueger, and Peter Steffen Schmidt. Esg rating disagreement and stock returns. *Financial Analysts Journal*, 77(4):104–127, 2021.

[122] Stefano Giglio, Bryan T Kelly, and Dacheng Xiu. Factor models, machine learning, and asset pricing. *Machine Learning, and Asset Pricing (October 15, 2021)*, 2021.

[123] Paolo Giudici, Branka Hadji-Misheva, and Alessandro Spelta. Network based credit risk models. *Quality Engineering*, 32(2):199–211, 2020.

[124] Paolo Giudici, Maura Mezzetti, and Pietro Muliere. Mixtures of products of dirichlet processes for variable selection in survival analysis. *Journal of statistical planning and inference*, 111(1-2):101–115, 2003.

[125] Paolo Giudici and Emanuela Raffinetti. Shapley-lorenz explainable artificial intelligence. *Expert systems with applications*, 167:114104, 2021.

[126] Florian Glaser and Luis Bezzenberger. Beyond cryptocurrencies-a taxonomy of decentralized consensus systems. In *23rd European conference on information systems (ECIS), Münster, Germany*, 2015.

[127] Periklis Gogas and Theophilos Papadimitriou. Machine learning in economics and finance. *Computational Economics*, 57(1):1–4, 2021.

[128] Itay Goldstein, Chester S Spatt, and Mao Ye. Big Data in Finance. *The Review of Financial Studies*, 34(7):3213–3225, 04 2021.

[129] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Vice: visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 531–535, 2020.

[130] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[131] M. J. Gordon. Dividends, earnings, and stock prices. *The Review of Economics and Statistics*, 41(2):99–105, 1959.

[132] Jerry R Green and John B Shoven. The effects of interest rates on mortgage prepayments, 1983.

[133] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[134] Branka Hadji Misheva, Ali Hirsa, Joerg Osterrieder, Onkar Kulkarni, and Stephen Fung Lin. Explainable ai in credit risk management. *Credit Risk Management (March 1, 2021)*, 2021.

[135] Kathleen Weiss Hanley and Gerard Hoberg. The information content of ipo prospectuses. *The Review of Financial Studies*, 23(7):2821–2864, 2010.

[136] Campbell R. Harvey, Yan Liu, and Heqing Zhu. … and the Cross-Section of Expected Returns. *The Review of Financial Studies*, 29(1):5–68, 10 2015.

[137] Md Hasan, József Popp, Judit Oláh, et al. Current landscape and influence of big data on finance. *Journal of Big Data*, 7(1):1–17, 2020.

[138] Michael Hay, David W Thomas, John L Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1):40–51, 2014.

[139] Alexander D Healy and Andrew W Lo. Managing real-time risks and returns: the thomson reuters newsscope event indices. *The Handbook of News Analytics in Finance*, pages 73–108, 2011.

[140] Donald Olding Hebb. Drives and the cns (conceptual nervous system). *Psychological review*, 62(4):243, 1955.

[141] Elaine Henry. Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, 45(4):363–407, 2008.

[142] Steven L. Heston and K. Geert Rouwenhorst. Industry and country effects in international stock returns. *The Journal of Portfolio Management*, 21(3):53–58, 1995.

[143] Gerard Hoberg and Gordon Phillips. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465, 2016.

[144] Allen Huang, Hui Wang, and Yi Yang. Finbert—a large language model approach to extracting information from financial text. *Available at SSRN 3910214*, 2020.

[145] Ethan J Hugstad. Biotech fever: Market overreaction to fda clinical trials. Technical report, University of Northern Iowa, 2018.

[146] Thomas J Hwang. Stock market returns and clinical trial results of investigational compounds: an event study analysis of large biopharmaceutical companies. *PloS One*, 8(8):e71966, 2013.

[147] Alice M. Isen. An influence of positive affect on decision making in complex situations: Theoretical issues with practical implications. *Journal of Consumer Psychology*, 11(2), 2001.

[148] Jinglin Jiang, Li Liao, Xi Lu, Zhengwei Wang, and Hongyu Xiang. Deciphering big data in consumer credit evaluation. *Journal of Empirical Finance*, 62:28–45, 2021.

[149] Needham Jones and Frank Bacon. Surprise earnings announcement: A test of market efficiency. In *Allied Academies International Conference. Academy of Accounting and Financial Studies. Proceedings*, volume 12, page 43. Citeseer, 2007.

[150] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 427–431, 2016.

[151] Narayanan Kandasamy, Ben Hardy, Lionel Page, Markus Schaffner, Johann Graggaber, Andrew S Powlson, Paul C Fletcher, Mark Gurnell, and John Coates. Cortisol shifts financial risk preferences. *Proceedings of the National Academy of Sciences*, 111(9):3608–3613, 2014.

[152] Christopher Kantos, Dan Joldzic, Gautam Mitra, and Kieu Thi Hoang. Comparative analysis of nlp approaches for earnings calls. *Available at SSRN 4210529*, 2022.

[153] Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. Predicting returns with text data. Technical report, National Bureau of Economic Research, 2019.

[154] Mozaffar Khan, George Serafeim, and Aaron Yoon. Corporate sustainability: First evidence on materiality. *The accounting review*, 91(6):1697–1724, 2016.

[155] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[156] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry investigation*, 15(3):235, 2018.

[157] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1746–1751, 2014.

[158] Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 423–430, 2003.

[159] Ozan Kose, Pinar Karagoz, and Gokce E Phillips. Towards a crypto asset taxonomy: a text classification-based approach. In *Proceedings of the 12th International Conference on Management of Digital EcoSystems*, pages 18–25, 2020.

[160] Sagar P Kothari and Jerold B Warner. Econometrics of event studies. In *Handbook of empirical corporate finance*, pages 3–36. Elsevier, 2007.

[161] Camelia M Kuhnen and Brian Knutson. The neural basis of financial risk taking. *Neuron*, 47(5):763–770, 2005.

[162] Neeraj Kumar, Stuart Doole, Ketaki Garg, Vishad Bhalodia, and Devika Ghate. Indexing change. *MSCI*, 2019.

[163] Azadeh Kushki, Jillian Fairley, Satyam Merja, Gillian King, and Tom Chau. Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. *Physiological measurement*, 32(10):1529, 2011.

[164] Håvard Kvamme, Nikolai Sellereite, Kjersti Aas, and Steffen Sjursen. Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207–217, 2018.

[165] Jens Lausen. Regulating initial coin offerings? a taxonomy of crypto-assets. In *Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden*, 2019.

[166] Charles M.C. Lee, Paul Ma, and Charles C.Y. Wang. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, 116(2):410–431, 2015.

[167] Mi-hee Lee, Gyunghye Yang, H-K Lee, and Seokwon Bang. Development stress monitoring system based on personal digital assistant (pda). In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 2364–2367. IEEE, 2004.

[168] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[169] Jennifer S. Lerner and Dacher Keltner. Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81(1):146—-159, 2001.

[170] Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. Emotion and decision making. *Annual Review of Psychology*, 66:799–823, 2015.

[171] Jennifer S. Lerner, Ye Li, and Elke U. Weber. The financial costs of sadness. *Psychological science*, 24(1):72—-79, 2014.

[172] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[173] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[174] Karl V Lins, Henri Servaes, and Ane Tamayo. Social capital, trust, and firm performance: The value of corporate social responsibility during the financial crisis. *the Journal of Finance*, 72(4):1785–1824, 2017.

[175] Jiatao Liu, Ian W Marsh, Paolo Mazza, and Mikael Petitjean. Factor structure in cryptocurrency returns and volatility. *Available at SSRN 3389152*, 2019.

[176] Weiyi Liu, Xuan Liang, and Guowei Cui. Common risk factors in the returns on cryptocurrencies. *Economic Modelling*, 86:299–305, 2020.

[177] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[178] Yukun Liu, Aleh Tsyvinski, and Xi Wu. Common risk factors in cryptocurrency. *The Journal of Finance*, 77(2):1133–1177, 2022.

[179] Andrew W. Lo. The adaptive markets hypothesis. *The Journal of Portfolio Management*, 30(5):15–29, 2004.

[180] Andrew W Lo, Carole Ho, Jayna Cummings, and Kenneth S Kosik. Parallel discovery of alzheimer's therapeutics. *Science translational medicine*, 6(241):241cm5–241cm5, 2014.

[181] Andrew W Lo and A Craig MacKinlay. Data-snooping biases in tests of financial asset pricing models. *The Review of Financial Studies*, 3(3):431–467, 1990.

[182] Andrew W Lo and Dmitry V Repin. The psychophysiology of real-time financial risk processing. *Journal of cognitive neuroscience*, 14(3):323–339, 2002.

[183] Andrew W Lo, Dmitry V Repin, and Brett N Steenbarger. Fear and greed in financial markets: A clinical study of day-traders. *American Economic Review*, 95(2):352–359, 2005.

[184] Andrew W Lo and Kien Wei Siah. Financing correlated drug development projects. *The Journal of Structured Finance*, 27(1):17–34, 2021.

[185] Andrew W Lo, Lan Wu, Ruixun Zhang, and Chaoyi Zhao. Optimal impact portfolios with general dependence and marginals. Available at SSRN 4177277, 2022.

[186] Andrew W Lo and Ruixun Zhang. Quantifying the impact of impact investing. *Available at SSRN 3944367*, 2021.

[187] George Loewenstein and Jennifer S. Lerner. The role of affect in decision making. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, editors, *Handbook of affective sciences*, pages 619–642. Oxford University Press, 2003.

[188] Alejandro Lopez-Lira. Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*, 2023.

[189] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.

[190] Tim Loughran and Bill McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016.

[191] Rafael Lozano, Nancy Fullman, John Everett Mumford, Megan Knight, Celine M Barthelemy, and Cristiana Abbafati. Measuring universal health coverage based on an index of effective coverage of health services in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 396(10258):1250–1284, 2020.

[192] Lorenzo Lucchini, Laura Alessandretti, Bruno Lepri, Angela Gallo, and Andrea Baronchelli. From code to market: Network of developers and correlated returns of cryptocurrencies. *Science Advances*, 6(51):eabd2204, 2020.

[193] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[194] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[195] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[196] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[197] Xiaojun Ma, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, and Xueqi Niu. Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31:24–39, 2018.

[198] A Craig MacKinlay. Event studies in economics and finance. *Journal of economic literature*, 35(1):13–39, 1997.

[199] Christopher R Madan, Elliot A Ludvig, and Marcia L Spetch. The role of memory in distinguishing risky decisions from experience and description. *The Quarterly Journal of Experimental Psychology*, 70(10):2048–2059, 2017.

[200] Prasanta Chandra Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

[201] Igor Makarov and Antoinette Schoar. Blockchain analysis of the bitcoin market. Technical report, National Bureau of Economic Research, 2021.

[202] Igor Makarov and Antoinette Schoar. Cryptocurrencies and decentralized finance (defi). Technical report, National Bureau of Economic Research, 2022.

[203] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

[204] Rosario N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

[205] Agnese Mariotti. The effects of chronic stress on health: new insights into the molecular mechanisms of brain–body communication. *Future science OA*, 1(3), 2015.

[206] David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, 183(3):1466–1476, 2007.

[207] Bruce S McEwen, Nicole P Bowles, Jason D Gray, Matthew N Hill, Richard G Hunter, Ilia N Karatsoreos, and Carla Nasca. Mechanisms of stress in the brain. *Nature neuroscience*, 18(10):1353–1363, 2015.

[208] William L Megginson, Angela Morgan, and Lance Nail. The determinants of positive long-term performance in strategic mergers: Corporate focus and cash. *Journal of Banking & Finance*, 28(3):523–552, 2004.

[209] MSCI Data Methodology. Global industry classification standard (GICS®) methodology: Guiding principles and methodology for gics. 2020.

[210] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

[211] Ewa A Miendlarzewska, Michael Kometer, and Kerstin Preuschoff. Neurofinance. *Organizational Research Methods*, 22(1):196–222, 2019.

[212] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[213] Nicholas Moehle, Stephen Boyd, and Andrew Ang. Portfolio performance attribution via shapley value. *arXiv preprint arXiv:2102.05799*, 2021.

[214] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[215] Thomas J Moore, James Heyward, Gerard Anderson, and G Caleb Alexander. Variation in the estimated costs of pivotal clinical benefit trials supporting the us approval of new therapeutic agents, 2015–2017: a cross-sectional study. *BMJ open*, 10(6):e038863, 2020.

[216] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[217] MSCI. MSCI peer similarity scores overview. *MSCI Inc.*, 2021.

[218] Stephen Muggleton. Inductive logic programming. *New generation computing*, 8(4):295–318, 1991.

[219] Christopher JL Murray, Aleksandr Y Aravkin, Peng Zheng, Cristiana Abbafati, Kaja M Abbas, Mohsen Abbasi-Kangevari, Foad Abd-Allah, Ahmed Abdelalim, Mohammad Abdollahi, Ibrahim Abdollahpour, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet*, 396(10258):1223–1249, 2020.

[220] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[221] Andreas Neuhierl, Anna Scherbina, and Bernd Schiusene. Market reaction to corporate press releases. *The Journal of Financial and Quantitative Analysis*, 48(4):1207–1240, 2013.

[222] John R Nofsinger, Fernando M Patterson, and Corey A Shank. Decision-making, financial risk aversion, and behavioral biases: The role of testosterone and stress. *Economics & Human Biology*, 29:1–16, 2018.

[223] Lennart Obermann and Stephan Waack. Interpretable multiclass models for corporate credit rating capable of expressing doubt. *Frontiers in Applied Mathematics and Statistics*, 2:1–16, 2016.

[224] Christopher B Overgaard, RA van den Broek, Jay H Kim, and Allan S Detsky. Biotechnology stock prices before public announcements: evidence of insider trading? *Journal of investigative medicine: the official publication of the American Federation for Clinical Research*, 48(2):118–124, 2000.

[225] Laura E Panattoni. The effect of paragraph iv decisions and generic entry before patent expiration on brand pharmaceutical firms. *Journal of health economics*, 30(1):126–145, 2011.

[226] L'uboš Pástor and Robert F Stambaugh. Costs of equity capital and model mispricing. *The Journal of Finance*, 54(1):67–121, 1999.

[227] L'uboš Pástor, Robert F Stambaugh, and Lucian A Taylor. Dissecting green returns. *Journal of Financial Economics*, 146(2):403–424, 2022.

[228] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature reviews Drug discovery*, 9(3):203–214, 2010.

[229] Ioannis Pavlidis, Panagiotis Tsiamyrtzis, Dvijesh Shastri, Avinash Wesley, Yan Zhou, Peggy Lindner, Pradeep Buddharaju, Rohan Joseph, Anitha Mandapati, Brian Dunkin, et al. Fast by nature-how stress patterns define human experience and performance in dexterous tasks. *Scientific Reports*, 2(1):1–9, 2012.

[230] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[231] W. J. Plath. Request: A natural language question-answering system. *IBM Journal of Research and Development*, 20(4):326–335, 1976.

[232] Ksenia Ponomareva and Simone Caenazzo. Interpretability of neural networks: A credit card default model example. *Available at SSRN 3519142*, 2019.

[233] Daniel Porath and Christian Schaefer. Applying the bass model to pharmaceuticals in emerging markets. *International Journal of Market Research*, 56(4):513–530, 2014.

[234] Lynnette Purda and David Skillicorn. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research*, 32(3):1193–1223, 2015.

[235] Ayoub El Qadi, Natalia Diaz-Rodriguez, Maria Trocan, and Thomas Frossard. Explaining credit risk scoring through feature contribution alignment with expert risk analysts. *arXiv preprint arXiv:2103.08359*, 2021.

[236] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[237] Thomson Reuters. Cmr international pharmaceutical r&d factbook. *Thomson Reuters, London*, 2012.

[238] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[239] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[240] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[241] Md Fahim Rizwan, Rayed Farhad, Farhan Mashuk, Fakhrul Islam, and Mohammad Hasan Imam. Design of a biosignal based stress detection system using machine learning techniques. In *2019 international conference on robotics, electrical and signal processing techniques (ICREST)*, pages 364–368. IEEE, 2019.

[242] J. R. Ross. *Constraints on variables in syntax*. PhD thesis, MIT, 1967.

[243] Jeffrey M Rothenstein, George Tomlinson, Ian F Tannock, and Allan S Detsky. Company stock prices before and after public announcements related to oncology drugs. *Journal of the National Cancer Institute*, 103(20):1507–1512, 2011.

[244] Cynthia Rudin and Yaron Shaposhnik. Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation. *Available at SSRN 3395422*, 2019.

[245] Anoop S Kumar and Taufeeq Ajaz. Co-movement in crypto-currency markets: evidences from wavelet analysis. *Financial Innovation*, 5(1):1–17, 2019.

[246] Apaar Sadhwani, Kay Giesecke, and Justin Sirignano. Deep learning for mortgage risk. *Journal of Financial Econometrics*, 19(2):313–368, 2021.

[247] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424, 2021.

[248] Christophe Schinckus, Dang Pham Thien Duy, and Nguyen Phuc Canh. Interdependences between cryptocurrencies: A network analysis from 2013 to 2018. *Journal of Interdisciplinary Economics*, 33(2):190–199, 2021.

[249] Norbert Schwarz and Gerald L. Clore. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3):513–523, 1983.

[250] Natalia Semenova and Lars G Hassel. On the validity of environmental performance metrics. *Journal of Business Ethics*, 132(2):249–258, 2015.

[251] Aylin Sertkaya, Hui-Hsing Wong, Amber Jessup, and Trinidad Beleche. Key cost drivers of pharmaceutical clinical trials in the united states. *Clinical Trials*, 13(2):117–126, 2016.

[252] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on information technology in biomedicine*, 14(2):410–417, 2009.

[253] Jaime Sevilla, Pablo Villalobos, Juan Felipe Cerón, Matthew Burtell, Lennart Heim, Amogh B Nanjajjar, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, Jean-Stanislas Denain, and Owen Dudney. Parameter, compute and data trends in machine learning. https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OIP 2022.

[254] Anurag Sharma and Nelson Lacey. Linking product development outcomes to market valuation of the firm: The case of the us pharmaceutical industry. *Journal of Product Innovation Management*, 21(5):297–308, 2004.

[255] William F Sharpe. Efficient capital markets: a review of theory and empirical work: discussion. *The Journal of Finance*, 25(2):418–420, 1970.

[256] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, 2013.

[257] Kien Wei Siah, Dermot P Maher, and Andrew W Lo. Financing pharmaceuticals and medical devices for pain treatment and opioid use disorder. *Emerging Trends in Drugs, Addictions, and Health*, page 100045, 2022.

[258] Kien Wei Siah, Qingyang Xu, Kirk Tanner, Olga Futer, John J Frishkopf, and Andrew W Lo. Accelerating glioblastoma therapeutics via venture philanthropy. *Drug Discovery Today*, 26(7):1744–1749, 2021.

[259] Naeem Siddiqi. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons, 2012.

[260] Justin Sirignano and Kay Giesecke. Risk analysis for large pools of loans. *Management Science*, 65(1):107–121, 2019.

[261] Alec Smith, Terry Lohrenz, Justin King, P Read Montague, and Colin F Camerer. Irrational exuberance and neural crash warning signals during endogenous experimental market bubbles. *Proceedings of the National Academy of Sciences*, 111(29):10503–10508, 2014.

[262] Bruno Solnik and Jacques Roulet. Dispersion as cross-sectional correlation. *Financial Analysts Journal*, 56(1):54–61, 2000.

[263] Kathryn Sprague. Innovative announcement effects in the pharmaceutical industry via stock market volatility. *University of Colorado, Boulder*, 957, 2015.

[264] T. Sprenger and Isabell M. Welpe. Tweets and peers: Defining industry groups and strategic peers based on investor perceptions of stocks on twitter. *eBusiness & eCommerce eJournal*, 2011.

[265] Richard Stanton and Nancy Wallace. The bear's lair: Index credit default swaps and the subprime mortgage crisis. *The Review of Financial Studies*, 24(10):3250–3280, 2011.

[266] Andreas Sturm, Michael J Dowling, and Klaus Röder. Fda drug approvals: time is money! *Journal of Entrepreneurial Finance, JEF*, 12(2):23–53, 2007.

[267] Faris Sumadi. Event-based biotechnology stock price movement: Valuing success and failure in biotechnology product development. Technical report, University of Massachusetts Boston, 2016.

[268] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[269] Prasanna Tantri. Fintech for the poor: Financial intermediation without discrimination. *Review of Finance*, 25(2):561–593, 2021.

[270] Paolo Tasca and Claudio J Tessone. Taxonomy of blockchain technologies. principles of identification and classification. *arXiv preprint arXiv:1708.04872*, 2017.

[271] Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168, 2007.

[272] Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The journal of finance*, 63(3):1437–1467, 2008.

[273] Richard T Thakor, Nicholas Anaya, Yuwei Zhang, Christian Vilanilam, Kien Wei Siah, Chi Heem Wong, and Andrew W Lo. Just how good an investment is the biopharmaceutical sector? *Nature biotechnology*, 35(12):1149–1157, 2017.

[274] David W Thomas, Justin Burns, John Audette, Adam Carroll, Corey Dow-Hygelund, and Michael Hay. Clinical development success rates 2006–2015. *BIO Industry Analysis*, 1(16):25, 2016.

[275] Lyn C Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172, 2000.

[276] Jun Tu and Guofu Zhou. Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies. *Journal of Financial Economics*, 99(1):204–215, 2011.

[277] Peter D Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*, 2002.

[278] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[279] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[280] Maximilian AM Vermorken. Gics or icb, how different is similar? *Journal of Asset Management*, 12(1):30–44, 2011.

[281] Christiaan H Vinkers, Renske Penning, Juliane Hellhammer, Joris C Verster, John HGM Klaessens, Berend Olivier, and Cor J Kalkman. The effect of stress on core and peripheral body temperature in humans. *Stress*, 16(5):520–530, 2013.

[282] Mark Weber, Mikhail Yurochkin, Sherif Botros, and Vanio Markov. Black loans matter: Distributionally robust fairness for fighting subgroup discrimination, 2020.

[283] Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.

[284] Dennis Wei, Sanjeeb Dash, Tian Gao, and Oktay Gunluk. Generalized linear rule models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6687–6696. PMLR, 09–15 Jun 2019.

[285] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[286] Theresa Wilson, Janyce Wiebe, and Philipp Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[287] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimating clinical trial success rates and related parameters in oncology. *Available at SSRN 3355022*, 2019.

[288] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019.

[289] Ke Wu, Spencer Wheatley, and Didier Sornette. Classification of cryptocurrency coins and tokens by the dynamics of their market capitalizations. *Royal Society open science*, 5(9):180381, 2018.

[290] Beat Wuthrich, Vincent Cho, Steven Leung, D Permunetilleke, K Sankaran, and Jian Zhang. Daily stock market forecast from textual web data. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, volume 3, pages 2720–2725. IEEE, 1998.

[291] NZ Xie, Lionel Page, DA Granger, and John M Coates. The biology of 'risk-on'. decreasing inflammatory and stress responses on a london trading floor. *bioRxiv*, page 473157, 2018.

[292] Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

[293] Xiaoman Xing, Zhimin Ma, Mingyou Zhang, Ying Zhou, Wenfei Dong, and Mingxuan Song. An unobtrusive and calibration-free blood pressure estimation method using photoplethysmography and biometrics. *Scientific reports*, 9(1):1–8, 2019.

[294] Jin Yong Yoo and Yanjun Qi. Towards improving adversarial training of nlp models. *arXiv preprint arXiv:2109.00544*, 2021.

[295] H. Zhang and G. Bonne. Peering into peer selection: Quantifying company similarity. *MSCI Research Insight*, 2019.

[296] H. Zhang and G. Bonne. Looking for a better hedge. *MSCI Research Insight*, 2020.

[297] Jiansong Zhang and Nora M El-Gohary. Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking. *Journal of Computing in Civil Engineering*, 30(2):04015014, 2016.

[298] Lydia Chuyue Zhang. Event study: The influence of covid-19 research and development news on pharmaceutical and biotech companies. 2021.

[299] Yanzhao Zou and Dorien Herremans. A multimodal model with twitter finbert embeddings for extreme price movement prediction of bitcoin. *arXiv preprint arXiv:2206.00648*, 2022.

[300] Ľuboš Pástor, Robert F. Stambaugh, and Lucian A. Taylor. Sustainable investing in equilibrium. *Journal of Financial Economics*, 142(2):550–571, 2021.