Applications of Machine Learning and First-Principle Modeling to Evaluate Design Enhancements in Autoinjectors
by
Ankita Singh

B.E. Chemical & Polymer Engineering, Birla Institute of Technology, Mesra, 2012
M.S. Plastics Engineering, University of Massachusetts, Lowell, 2014

Submitted to the Department of Mechanical Engineering and the MIT Sloan School of Management in the Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
and
Master of Science in Mechanical Engineering

In conjunction with the Leaders for Global Operations Program at the

Massachusetts Institute of Technology
June 2021

Signature of Author……
Department of Mechanical Engineering and Sloan School of Management
May 14, 2021

Certified by……
Ellen Roche, Thesis Supervisor
W.M. Keck Foundation Career Development Professor
Department of Mechanical Engineering

Certified by……
Jonas Oddur Jonasson, Thesis Supervisor
Class of 1943 Career Development Professor
MIT Sloan School of Management

Accepted by……...
Nicolas Hadjiconstantinou
Chair, Mechanical Engineering Committee on Graduate Students

Accepted by……...
Maura Herson
Assistant Dean, MBA Program
MIT Sloan School of Management

THIS PAGE INTENTIONALLY LEFT BLANK

Applications of Machine Learning and First-Principle Modeling to Evaluate Design Enhancements in Autoinjectors

by Ankita Singh

Submitted to the Department of Mechanical Engineering and the MIT Sloan School of Management on May 14, 2021, in the partial fulfillment of the requirements for the degrees of Master of Science in Mechanical Engineering and Master of Business Administration in conjunction with the Leaders for Global Operations program.

Abstract

One of the key guiding principles in Amgen operations is to ensure reliability of the offered combination products to best serve patients and maintain a competitive advantage for the business. With a broad device portfolio and increasing sales volume, Amgen now has access to large data repositories and an opportunity to realize its value to "Be Science Based" and utilize this data in innovative ways to improve product designs and training programs. The goal of this project is to use machine learning to augment design decisions and provide products that truly resonate with Amgen's mission - "To serve patients".

This thesis presents a hybrid machine learning and first principle based model that can be used by Amgen to enhance the feedback loop between user experience and product design teams. By leveraging data on predicate autoinjector devices, we created models that can produce user experience insights and provide predictive capabilities for future product designs.

The methodology to generate our models relies on theoretical first principle modeling and data science. We utilized domain knowledge to extract product attributes that contribute towards user experience. One such attribute was drug injection time for an autoinjector. The theoretical model used autoinjector design and drug product features to predict drug injection time. The machine learning model used drug injection time data along with other product design parameters to predict user experiences.

The results of our model provided a direct link between the design attributes and user feedback metric. The accuracy of the hybrid model varied between 50-70% depending on the algorithm used. The first principle model results were very close to the empirical injection time data with only 12% error.

Furthermore, the thesis presents an in-depth analysis on the interpretability of results by utilizing techniques like partial dependence and permutation variable importance charts to enhance the understanding of results generated by a machine learning model.

Thesis Supervisor: Ellen Roche, Thesis Supervisor
Title: W.M. Keck Foundation Career Development Professor
MIT Department of Mechanical Engineering/Institute for Medical Engineering and Science

Thesis Supervisor: Jonas Jonasson
Title: Class of 1943 Career Development Professor

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

I would like to express my gratitude to MIT's Leader for Global Operations program staff and students for their support during this project. This project would not have been successful without their continuous mentorship and encouragement.

I would like to thank my internship supervisor, Hillary Doucette, and the extended Amgen Integrated Product Surveillance team members for providing me all the information and support needed to succeed in this project. I would also thank our executive sponsor, Jennie Stevenson, for providing us this exciting opportunity to intern at Amgen and learn from this experience.

I would like to thank my thesis advisors, Prof. Jonas Jonasson and Prof. Ellen Roche for being patient with my questions and provided me valuable suggestions that helped me successfully achieve the project deliverables.

Last but not the least, I would like to thank my husband, Bhaskar, and my parents for being there for me through all the ups and downs and supported me during this challenging year.

Table of Contents

Table of Figures

Glossary/Acronyms

CAGR	Cumulative Annual Growth Rate
CART	Classification and Regression Trees
DNA	Deoxyribonucleic Acid
EDL	Enterprise Data Lake
EpiPen	Epinephrin Autoinjectors
FPT	Final Product Technologies
FY	Fiscal Year
GBDT	Gradient Boosting Decision Trees
GBM	Gradient Boosting Model
GMP	Good Manufacturing Practices
ICE	Individual Conditional Expectations
IJT	Injection Time
IVD	In Vitro Diagnostic Assay
OLS	Ordinary Least Squares
PDP	Partial Dependency Plot
QC	Quality Control
R&D	Research and Development
R&D	Research and Development
SHAP	Shapley Additive Explanations
SQL	Structured Query Language
U.S.	United States
UFM	User Feedback Metric
USFDA	United Stated Food and Drug Administration
V&V	Design Verification and Validation
XGBoost	Extreme Gradient Boosting

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1: Project Introduction

1.1 Project Motivation and Opportunity

Amgen's Final Product Technologies (FPT) organization focuses on the design, development, distribution, and post-market surveillance of Amgen's drug delivery devices, ranging from advanced electromechanical systems to conventional pre-filled syringes. Over the last five years, Amgen's portfolio of devices and combination products has vastly grown in volume and complexity, carving out opportunities to harness a wide scale of information in novel ways to improve product performance and usability.

Access to commercial data like post-market surveillance provides a mechanism for continuous feedback within the quality management system to enable improvement measures and monitor product/process risks. Furthermore, it is a vital source of information when launching a new product or assessing the impact of a design change on an on-market product.

With an increase in focus within Amgen on data strategies for devices and combination products, there is an opportunity to develop industry-leading analytical tools that evaluate operations and commercial data to translate insights into user-centric design enhancements for next-generation products.

1.2 Project Goals

The primary goal of the internship is to develop a hybrid model to directly correlate design inputs impacting drug injection time (from theoretical models) to user feedback. The business value-add of such a model will be a closed feedback loop be- tween design and post-market surveillance teams that will facilitate faster continuous improvement efforts to enhance user experiences. In support of this objective, the thesis seeks to address the following questions:

1. What is the impact of a change in design specifications on the drug injection parameter and user experience?
2. Can we utilize first principles in device and process modeling with reliable accuracy to predict functional attributes, like injection time, in an autoinjector?
3. How can we understand the partial dependency of a single variable in a machine learning model designed with multivariate principles in order to quantify its impact?

1.3 Statement of Hypothesis and Project Approach

There are two hypotheses of this project as mentioned below:

1. Changes in design specifications of an autoinjector can have a measurable impact on user experience.
2. First-principle modeling can be used reliably to generate input data (theoretical) for machine learning models built with empirical data (experimental).

The methodology to generate our models relied on theoretical first-principle modeling and predictive machine learning modeling. A machine learning model was developed using key device design and performance attributes such as injection time from predicate device data to predict user feedback. First-principle modeling was

utilized to generate a theoretical prediction of injection time based on design specifications. Thus, by using the above two models in conjunction, we created a hybrid model that provides a direct link between the design attributes and user feedback metric. Fig 1 shows the visual representation of model architecture.

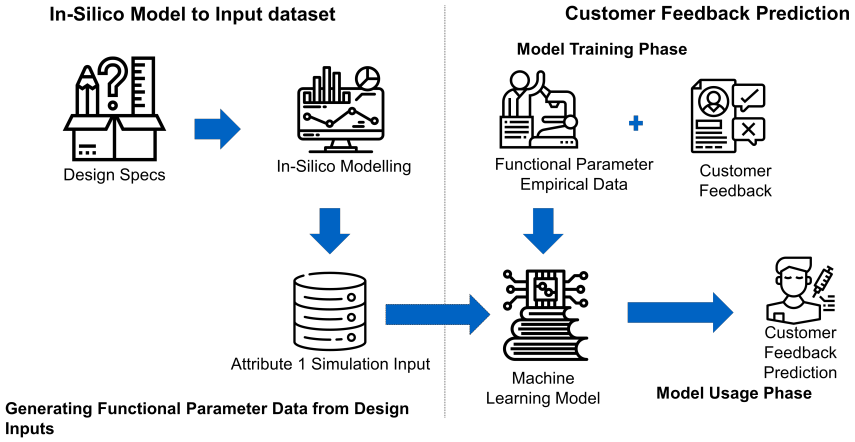

Figure 1: Visual representation of the hybrid model

We used the following parameters to evaluate the performance of our hybrid model:

1. **Predictive accuracy**: the models developed should enable us to reliably predict the impact of a design change on user experiences.
2. **Interpretability**: the models should generate comprehensible insights directly linking the design specifications to user experience.

The model performance was measured on out-of-sample data to quantify predictive power. To enable interpretability, we incorporated tools like partial dependency plots that provided an easily understandable visual representation of the input variable under consideration against the outcome i.e., user feedback.

1.4 Thesis Outline

The thesis is organized as follows:

Chapter 2 - Background contains three subsections which provides an overview of the Biopharmaceutical industry and Amgen, Combination Products and Product Complaint and Post Market Surveillance. The goal of this chapter is to introduce the frequently used terms to the reader with relevant information and background.

Chapter 3 - Literature Review provides an overview of the literature survey in this research topic. In this chapter, I reviewed the industry relevant standards and recent advancements.

Chapter 4 - Research Methodology provides an in-depth analysis of the research methodologies used for this project. This chapter will include the theoretical understanding of the first-principle and machine learning algorithms used. It will also include the step-by-step explanation of data acquisition and pre-processing techniques along with its limitations and assumptions.

Chapter 5 - Results & Conclusions provides detailed analysis of the model performance by using the techniques mentioned in Chapter 4. Furthermore, this chapter will also include an overview of the design variables and its corresponding importance in the overall predictive power of the model. With the use of case study examples based on predicate devices, readers can get an understanding of the step-by-step implementation process and the outcomes. It will contain highlights of the key insights generated. It will also provide the business impact of this project for Amgen and finally concludes the thesis with recommendations for future expansion of this project.

Chapter 2: Background

2.1 Biopharmaceuticals

The scope of this thesis covers combination medical device design process and its impact on user experiences. This section provides an overview of the biopharmaceutical industry and one of its major players: Amgen Inc.

Biopharmaceuticals are pharmaceuticals (medicinal products, therapeutics, prophylactics and *in vivo* diagnostics) with active agents inherently biological in nature and manufactured using biotech [1]. To further explain, a biopharmaceutical (also known as a biologic), is produced by genetically engineering living cells, or tissues extracted or semi-synthesized from biological sources like humans, animals, or microorganisms [2]. These cells are grown in sophisticated bioreactors and the protein developed is extracted, purified and processed as drugs.

Humulin, the first recombinant DNA human insulin, was developed in 1978 by David Goeddel and his colleagues (of Genentech) by utilizing and combining the insulin A- and B- chains expressed in *Escherichia coli* [3]. In 1982, Humulin received USFDA (United States Food and Drug Administration) approval and thus became the first biopharmaceutical drug marketed [4]. Before this, each diabetic in need of insulin was consuming approximately one pig per week as the source for pancreatic tissue, the raw material for insulin manufacturing [5]. Since 1982, biopharmaceuticals have seen rapid advancements and a large number of biopharmaceutical drugs are now available for clinical use. Figure 2 below shows the number of biopharmaceuticals approved by USFDA between 2000-2017.

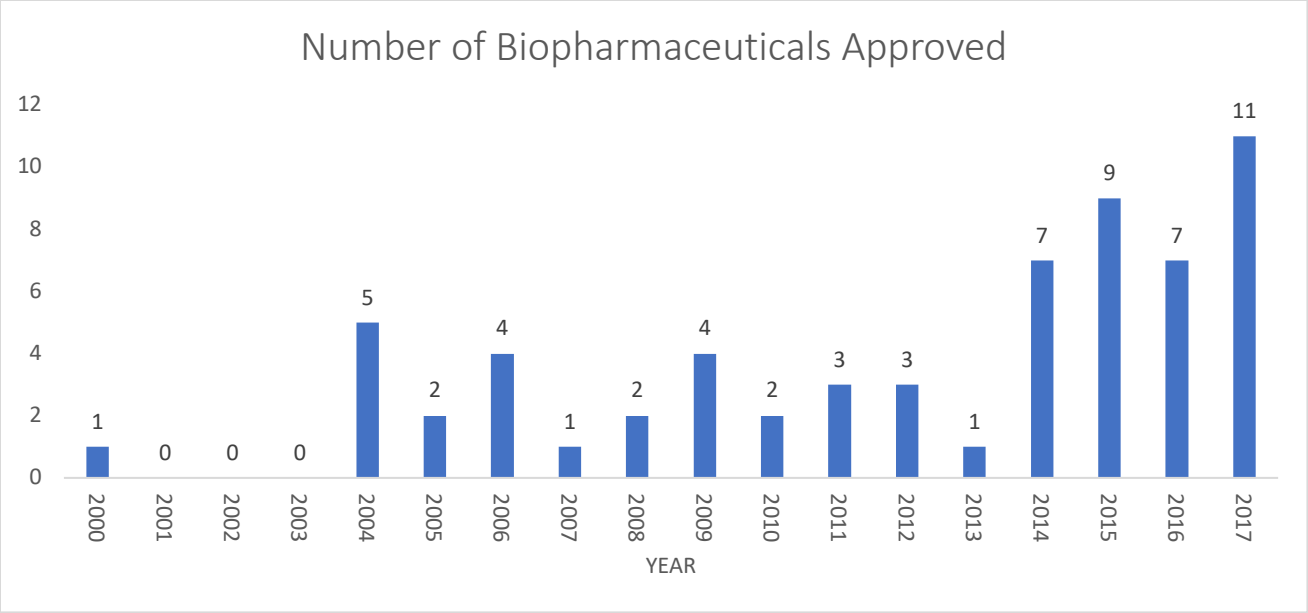

Figure 2: Number of biopharmaceuticals approved by USFDA between 2000 – 2017 [6]

Unlike chemical drugs, biopharmaceuticals development is a very time and resource-intensive process due to their extensive testing requirements. According to an article published in Biotech Healthcare, a typical manufacturing process for a chemical drug might contain 40 to 50 critical tests. The process for a biologic might contain 250 or more [7]. Additionally, biopharmaceuticals production uses specialized machinery and takes a disproportionately long time in the construction and validation of production facilities. These are some of the reasons which cause the cost differential between a chemical drug and a biopharmaceutical.

Even after these challenges in developing a biopharmaceutical product, the industry remains all time strong and continually improving in financial health. The global market for biopharmaceuticals is over $275 billion and is forecasted to grow at an 8.7% CAGR (Cumulative Annual Growth Rate) through 2025 [8]. New products and expanded approvals for new indications, as seen in Figure 2, are driving this market growth.

2.2 Amgen Inc.

Amgen (Applied Molecular Genetics Inc.) was established in Thousand Oaks, California, on April 8, 1980. In 1989, Amgen launched its first biopharmaceutical drug, Epogen, which later became one of the most successful drugs in the history of biopharmaceuticals [9].

Today, as one of the leading biopharmaceutical companies in the world, Amgen offers a wide portfolio of drugs for cardiovascular disease, bone disease, inflammation, oncology and nephrology. As of March 2020, Amgen employees approximately 22,000 people across its facilities worldwide. In 2019, Amgen earned a total revenue of $23.4 billion with $22.2 billion in product sales and a net Research and Development (R&D) spend of $4.0 billion. Currently, Amgen has 23 different products across the categories mentioned above [10].

2.3 Combination Products:

According to the USFDA, a combination product is a product composed of any combination of a drug and a device; a biological product and a device; a drug and a biological product; or a drug, device, and a biological product (FDA 21 CFR 3.2(e)).

Some examples of combination products can be seen in Fig 3.

Figure 3: Example of Combination Products (From Left to Right: On Body Drug Injector Device, Autoinjector, Inhaler)

Technological progress in drugs and delivery mechanisms has resulted in increased application of combination products in many environments. These products improve the patient experience by providing a convenient option to administer prescribed drug dosage without going to a hospital or a medical center.

As a combination product contains a combination of drug or biologic product and device, it becomes critical to create an overlapping development process. Figure 4 below gives a very good representation of one such co-development process.

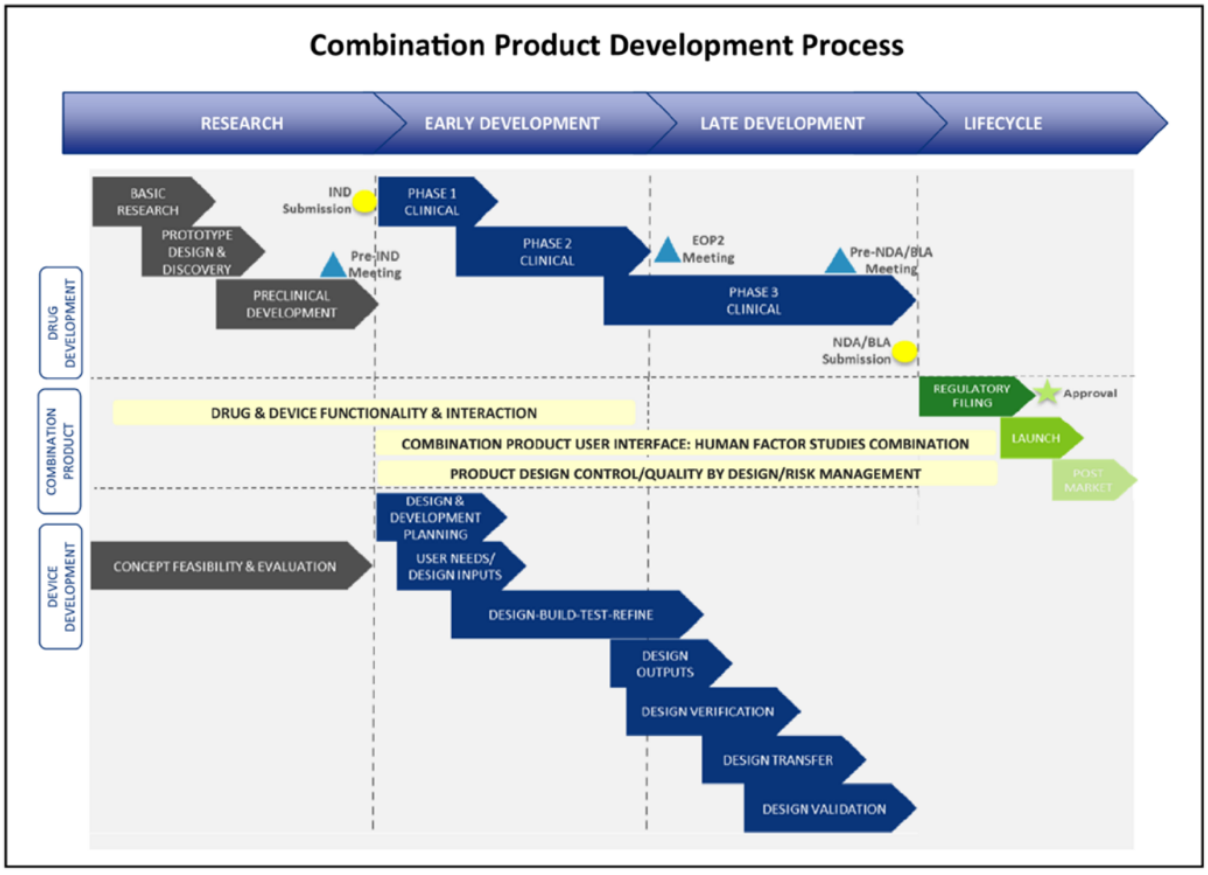

Figure 4: Combination Product Development Process [11]

2.3.1 Autoinjectors

Autoinjectors are devices designed to be used by the patient at home to self-administer drug dosage with minimum opportunities for error. These devices minimize or eliminate interactions with needles, removing a significant barrier to adherence for many patients. Needles are safely retracted or otherwise protected after a single use, eliminating the dangers of accidental needle sticks or potential reusing needles. These prefilled devices, when used correctly, guarantee correct dosage of medication [12]. Autoinjectors can be mechanical, electrical or a combination of both. Epinephrine autoinjectors (or EpiPen) are one of the most common autoinjectors. Fig 5 shows a basic construct of autoinjectors as a combination product.

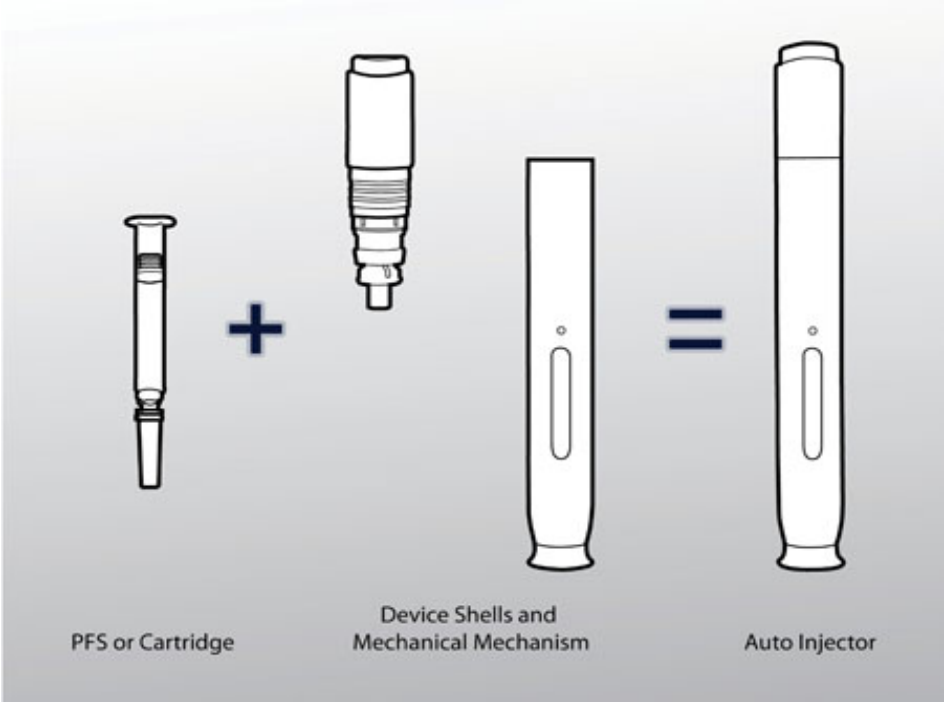

Figure 5: Autoinjector as a combination product *[13]*

2.4 Product Complaint & Post Market Surveillance

According to the USFDA, Complaint a complaint means any written, electronic, or oral communication that alleges deficiencies related to the identity, quality, durability, reliability,

safety, effectiveness, or performance of a device after it is released for distribution (FDA 21.CFR.820.3(b)).

In a combination product, a complaint might arise due to multiple reasons. Sometimes, a device that meets the requirements to ensure minimum risks before being introduced into the market may exhibit new issues in the post-market scenario. Due to these reasons, medical device manufacturers have an obligation to monitor their device performance through post market surveillance, and must have robust procedures in place to evaluate, investigate, and address complaints where required. This post market data provides a mechanism for continuous feedback within the quality management system to necessitate improvement measures and monitor product/process risks. Furthermore, it is a vital source of information when launching a new product or assessing the impact of a design change to an on-market product.

As Ogrodnik (2013) describes, post market surveillance captures information about a device that can be used in a three-pronged approach. The first prong adds to the clinical literature and knowledge base. The second prong provides marketing and sales information. The third prong provides quality related information and material from the technical knowledge base [14]. Figure 6 below showcases the components these three prongs from a post market surveillance in details.

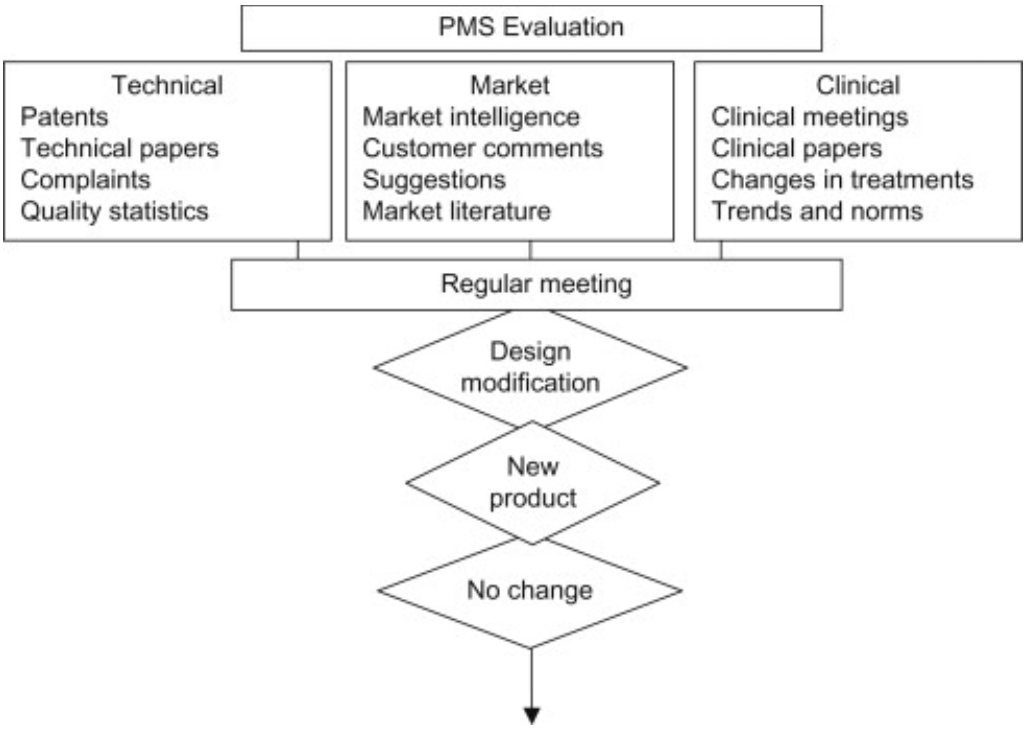

Figure 6: Post Market Surveillance Evaluation Process [14]

One of the key aspects of this project was to enhance the feedback loop between post-market surveillance data for autoinjectors to support evaluation of product design decisions related to drug injection.

Chapter 3: Literature Review

3.1 Medical Device Design Control Process

During the design and development of a medical device product, the manufacturers are required to ensure that the device is both safe and effective for users whose health may be on the line. Regulatory agencies like USFDA regulate this process and ensure that these devices are developed under Design Control procedures. According to the USFDA, Design controls are an interrelated set of practices and procedures that are incorporated into the design and development process, i.e., a system of checks and balances. [15]

Design controls (21 CFR 820.30) apply to all class II, class III and a limited number of class I medical devices.

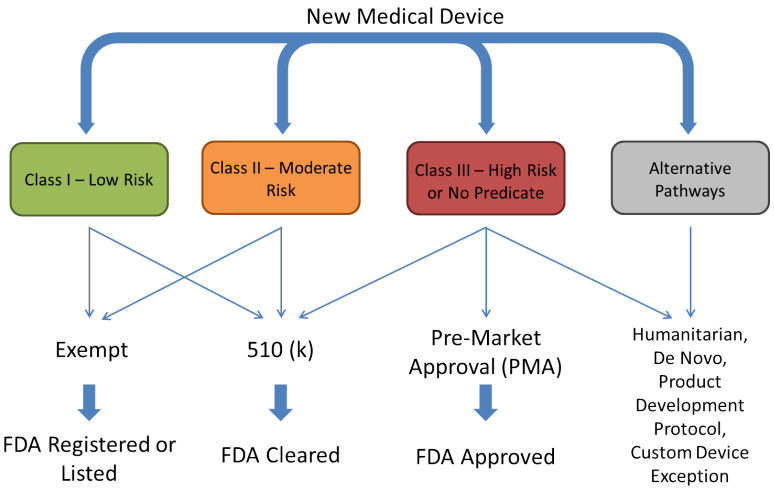

Figure 7: Regulatory Process Classification for Different Classes of Medical Devices [16]

A biopharmaceutical-based combination product development process starts with the end-user considerations. End-users' interaction with the device is the primary consideration in designing the device delivery mechanisms. For example, the development process of a device for self-administration, i.e., minimal or no supervision needed, would be different from those monitored by

health care practitioners. This process is iterative and includes several preliminary lab characterizations and human factors trials to refine the specifications. After completing the initial design process, a set of design outputs is then tested against the input criteria, and the product is ultimately validated against the original user needs. The design process also incorporates risk assessment methodologies as instructed in ISO 14971.

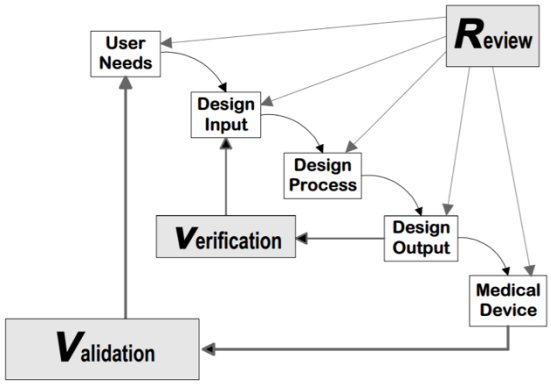

Figure 8: Application of Design Controls to the Waterfall Design Process

Figure 9: Different phases in the medical device commercial development process [15]

As the device moves from the design to the manufacturing phase, the production process is optimized and documented under the GMP or Good Manufacturing Practices guidelines as per the USFDA regulations. The production units are then taken for Design Verification and Validation (V&V) testing. The V&V testing ensures that the production unit meets the desired product specification through multiple lab-based and clinical trials. It is critical to incorporate appropriate statistical sample sizes

when conducting the V&V studies for process and design validation. This will also be based on stable manufacturing variability assumptions.

Design control practices apply to new product launches and continue after design transfer. When a design change is proposed for an on-market product, it must undergo appropriate levels of assessment, including validation, and where appropriate, verification, review, and approval before implementation. The level of assessment required shall be commensurate with the significance of the change and is processed through the manufacturer's change control program.

3.2 First Principle Modeling of Autoinjectors

The first principle modeling of combination devices like autoinjectors is studied extensively to understand the physical and mechanical system's drug flow properties. These first principle-based simulation models are very useful in calculating the device's key performance parameters in the early stages of product development. Product design engineers and drug development chemists often use these simulation results to finetune the drug product characteristics like viscosity, density, etc. The corresponding device design characteristics like needle dimensions, spring force, etc. to meet the desired product performance specs. In this section, we will discuss the first principle models of an autoinjector device. Figure 10 below shows the mechanical construct of an autoinjector device:

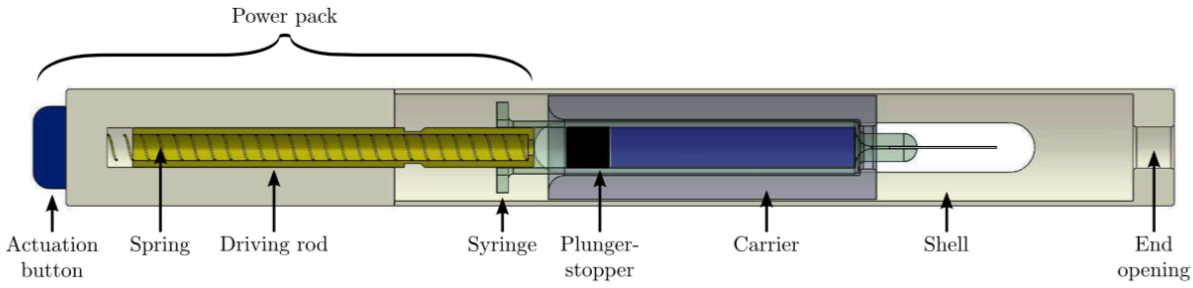

Figure 10: Mechanical Construct of an Autoinjector Device [17]

After simplification, the above autoinjector can be shown by Figure 11 below.

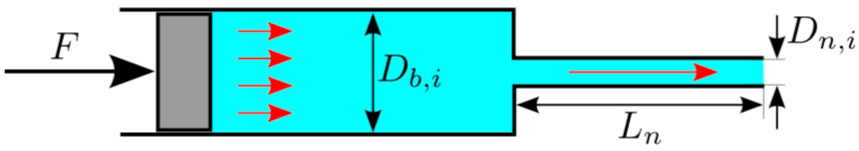

Figure 11: Simplified Flow Diagram of an Autoinjector [17]

Here D_b represents the diameter of syringe barrel, D_n represents the diameter of the needle, L_n represents the length of needle and μ is the viscosity of drug product.

By first principles, the force (F) required to inject a drug volume V in time t can be calculated by solving the Hagen-Poiseuille fluid flow equation [18]. The relationship between force, F, also known as injection force and other device parameters can be seen in equation 1 below:

$$F = 32\,\mu\,L_n\left(\frac{D_b^2}{D_n^4}\right)\left(\frac{V}{t}\right) \qquad \text{(Equation 1)}$$

Equation 1 provides a simple model to calculate some key parameters like injection time (t) given the injection force (F) and vice versa, which is an excellent first step during the device design process. However, this equation also has its limitation as it fails to incorporate some of the key device usage assumptions. In the scenario above, the simulation model calculates all the design parameters for injection in air and ignores the resistance and pressure gradience in the subcutaneous tissue, which is what the actual case will be during an injection administration. Figure 12 below shows an alternative representation of the drug injection force balance by separating the fluid force, frictional force, and the plunger force.

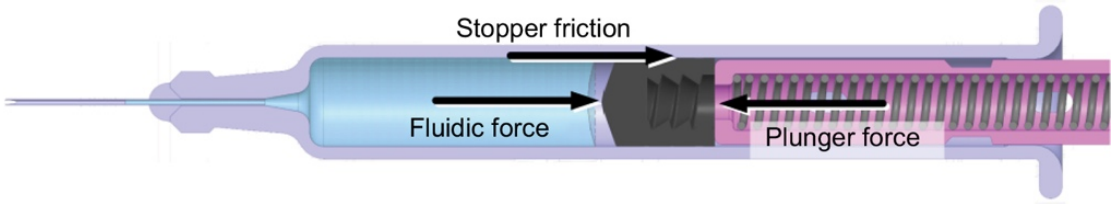

Figure 12: Equilibrium Forces in an Autoinjector Assembly *[19]*

$$F_{plunger} = F_{friction} + F_{fluidic} \qquad \text{(Equation 2)}$$

Combining equation 1 and 2, we can rewrite the injection time as a function below:

$$t_{inj} = f(\mu, L_N, R_B, R_N^4, x, 1/(F_{plunger} - F_{friction})) \quad \text{(Equation 3)}$$

where t_{inj} is injection time, μ is viscosity, L_N is needle length, x is stopper displacement, R_B is inner radius of syringe barrel, R_N is inner radius of needle, $F_{plunger}$ is plunger force, $F_{friction}$ is frictional force and $F_{fluidic}$ is fluid resistance force caused by the pressure drop in the needle.

The approach described above provides a good estimation for injection time using the design parameters of the autoinjector. However, it is still based on fundamental assumptions, such as the fluid under consideration (drug product), which is assumed to be a Newtonian fluid that maintains a constant viscosity with change in shear. This assumption can result in a discrepancy in the estimated and actual results.

In research by Thueer et al. (2018), we can see an attempt to model viscosity in terms of protein concentration and temperature. Figure 13 below shows the impact of change in viscosity on the injection time parameter.

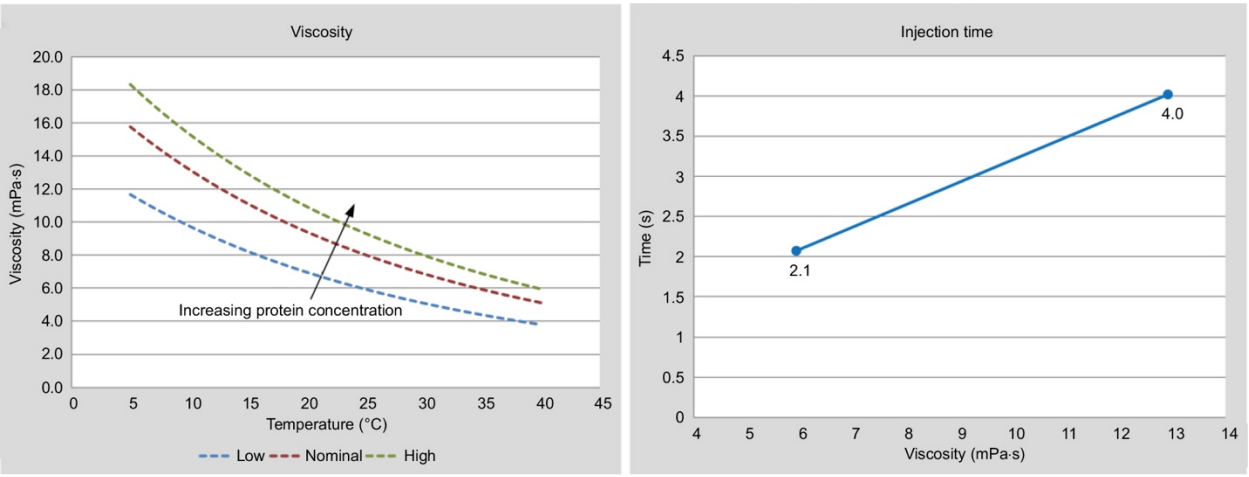

Figure 13: Impact of change in viscosity on injection time [19]

Other important research on the first principle modeling of autoinjectors was performed by Rathore et. al. (2011) in their research article on "Variability in syringe components and its impact on functionality of delivery systems". Rathore et.al. further developed the theoretical model in their

research and showed the impact of variability in syringe components on the device functionality. The frictional force in this paper was formulated for a siliconized surface which helps reduce the amount of friction between stopper and syringe body. Figure 14 below shows the syringe schematic with siliconized surface.

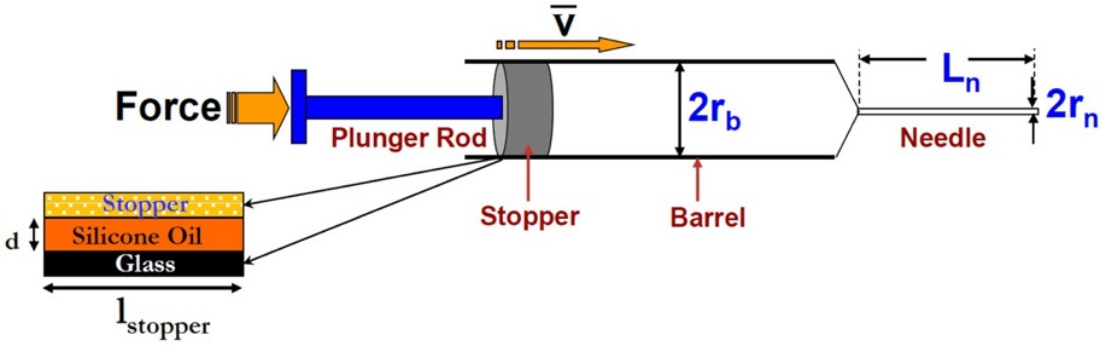

Figure 14: Syringe Force Schematic [20]

After including the fluid and friction forces, the plunger force can be derived as Equation 4 below.

$$F_{plunger} = \left(\frac{2\pi\mu_{oil} r_b l_{stopper}}{d_{oil}}\right)\bar{v} + \left(\frac{8\pi\mu l_n r_b^4}{r_n^4}\right)\bar{v} \qquad \text{(Equation 4)}$$

where μ_{oil} is the viscosity of lubricating oil, d_{oil} is the thickness of lubrication layer, $l_{stopper}$ is the length of the stopper in contact with glass, μ is the viscosity of drug product and $\bar{v}$ is the injection speed (linear piston speed with dimensions of length over time).

Figure 15 below shows the performance of the model built on equation 4.

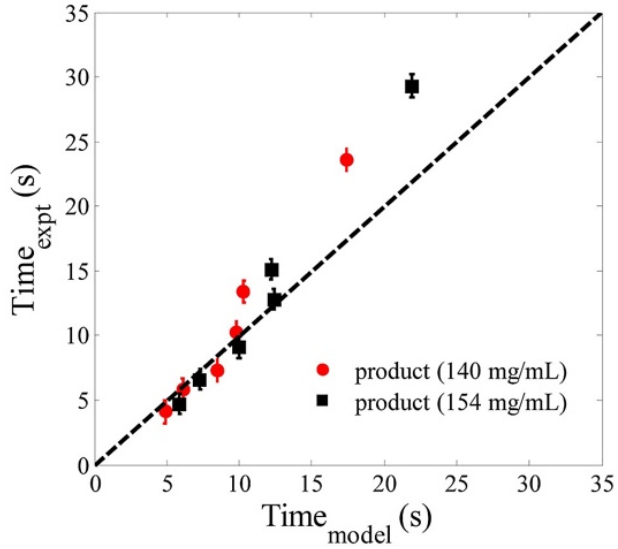

Figure 15: Injection Time - Model output vs Experimental for two different products [20]

3.3 Machine Learning and Big Data Analytics on Post-Market Surveillance Data

Using machine learning and big data analytics on user feedback data has gained popularity in recent years. With advancements in data acquisition and data processing techniques, businesses are now proactively tracking user behavior to stay ahead of the competition and continue to keep the market share. Service industries like banking and hotels leverage the learnings from user feedback to introduce highly desirable user features to the market. Another benefit of tracking user feedback is understanding the root cause of issues reported. Understanding these issues can help product design teams address them in future product releases.

Although significant progress has been made on using data science for user feedback analysis in industries like hotels, online retail, etc., the medical devices industry is still new in this research area. One of the significant works in this category comes from the University of Ottawa by Stephane Aris-Brosou et.al (2018). This work aims to integrate the quality control (QC) data with user complaints for In Vitro Diagnostics Assays (IVD). The machine learning techniques used for classification problems in this paper covered one simple classifier, CART (classification and regression trees), and

one sophisticated algorithm, adaptive boosting. The researchers were quite successful in their approach to predict the potential cause of user complaints based on the QC data with a classification error close to zero with the adaptive boosting algorithm. [21]

Chapter 4: Research Methodology

4.1 Contextualization

The scope of this thesis covers the combination medical device design process and its impact on user experience. This section provides an overview of the research methodology. We began this research project by first defining the problem. We divided the entire problem statement into several small segments and analyzed its technical and business impact. For technical impact, we aimed to create a reliable machine learning and first principle-based model to assist the design and development of future products. For business impact, this translated to achieving one of the key objectives at Amgen: improved user experience. The contextual visualization of our project approach in the form of smaller segments can be seen in Figure 16 below.

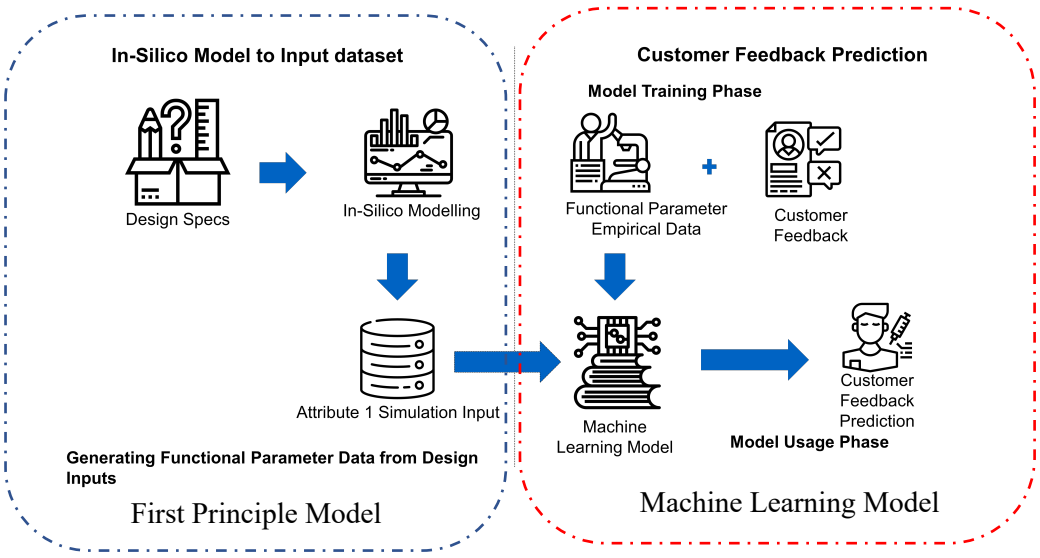

Figure 16: Technical Sections of the Project

4.2 Data Collection and Preprocessing

We used data querying from the enterprise data lake (EDL) for most user feedback and manufacturing batch data. Occasionally, we would receive special data upon request from suppliers in the form of excel sheets which was then cleaned and uploaded in the database for easier future query through EDL. We used Databricks as the primary interface for both data query and data analysis. Databricks provided an easier way to work collaboratively on this project. Its notebooks, an example seen in Figure 17, were capable of running multiple programming languages like Python, R, SQL etc. within the same page and hence enabled a lot of flexibility.

Figure 17: Databricks Notebook Example [22]

Data Preprocessing included several steps such as data cleaning, merging, imputing and removing outliers. Figure 18 shows a flow chart of the data merging process.

Figure 18: Dataframes merged to create final data table (Note that these tables are merged using a common identifier, Batch No. in this case.)

In order to keep the project work within scope, the data was analyzed for one of the product categories using a pre-selected autoinjector type. Further, the user feedback database was most comprehensive for US, thus the product manufacturing and supplier data was filtered to contain only batches that were sold in the US.

A large portion of the data requested for our analysis was not generated at Amgen, nor was it required for in process controls at Amgen. Thus, for select parameters, we relied on historical datasets from Amgen suppliers outside of existing quality agreements in place, leading to data availability constraints for select components and/or date ranges assessed. In some cases, the raw component measurements from suppliers were as low as 30% of the selected batches. This proportion of missing data required a reliable data imputation strategy. We used guidance from domain experts to determine the acceptable standard deviation for the missing supplier data. We then used a column-wise mean and standard deviation combination to impute missing values. This strategy was acceptable as the standard deviation was small enough to be considered insignificant for some of these features.

Further details about the data attributes and features will be explained in chapter 5.

4.3 First Principle Modelling

The first principle model can be defined as an engineering design model which is based on fundamental physical laws such as mass transfer, heat transfer, fluid flow, etc. These models require extensive domain knowledge to be able to model physical systems as equations. However, one of the key benefits of such a modeling technique is that it doesn't require empirical data for its development. Empirical data is often used to check the accuracy of the model and compensate for difficult-to-model errors. This technique of modeling is very frequently used in various chemical and biological process engineering. Compared to a data-driven model, first principle models offer more robustness with the ability to handle more complex and non-linear relations. It is also easy to interpret since it is fundamentally based on physical principles. One of the key challenges in this type of modeling is the need for deep domain expertise to understand and build such models. Depending on the complexity of the model, it can sometimes be very time-consuming.

Because of these limitations, first principle modeling is now used extensively in conjunction with empirical data and machine learning models. Such models can improve their accuracy as more data is collected, increasing its reliability. A 2 x 2 matrix, as seen in Figure 19, provides a relevant decision chart for modeling approaches based on domain knowledge and the volume of available data.

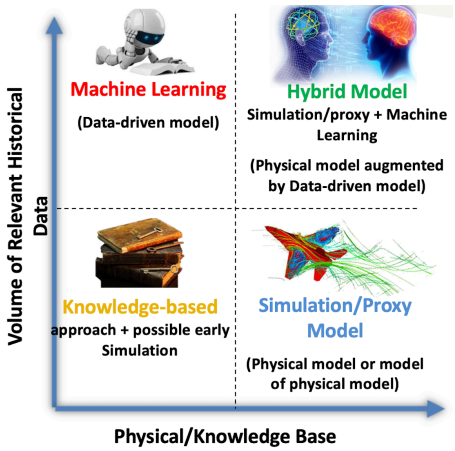

Figure 19: Hybrid Process Model *[23]*

In this project, we used the first principles in conjunction with empirical data driven modeling to build a hybrid model that can be used for future products and design changes without the need for empirical data.

4.4 Data Driven Modeling: Machine Learning

Machine learning is a method of data analysis that automates model building and decisions based on data. It is a branch of artificial intelligence based on the premise that algorithms can learn from data, identify patterns and make decisions with minimal human intervention. Machine-learning algorithms use statistics to find patterns in large data sets. Machine learning enables the analysis of extremely large datasets to gain insights and make better decisions in the future based on data.

Machine learning algorithms can be characterized into three broad categories:

- Supervised Learning Algorithms
- Unsupervised Learning Algorithms
- Reinforcement Learning Algorithms

Supervised Learning Algorithms:

In supervised learning, the training data is labeled to tell the algorithm the pattern it needs to learn. The aim of the algorithm is to develop a function that maps input to output. Once the algorithm is trained, it can predict the output based on new data (inputs). Popular techniques for supervised learning are decision trees, neural networks, linear regression, logistic regression etc.

Unsupervised Learning algorithms:

In unsupervised learning, the training data has no labels, and the algorithm tries to find the pattern in the data on its own. The aim of the algorithm is to find a function that describes the hidden structure in the training data. Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering, and singular value decomposition.

Reinforcement Learning algorithms:

Reinforced learning tries to achieve a clear objective based on trial and error. This is often used in dynamic environments where a software agent interacts with the environment to accomplish specific tasks, e.g., playing a game against an opponent.

In this project, we will be working with labeled datasets and hence use supervised learning. We will be using XGboost, Gradient Boosting Models, Random Forest, and Ordinary Least square regression to perform supervised learning.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way [24]. XGBoost is used for supervised learning problems, where we use the training data (with multiple features) to predict a target variable. Figure 20 below shows the architecture of a XGBoost model.

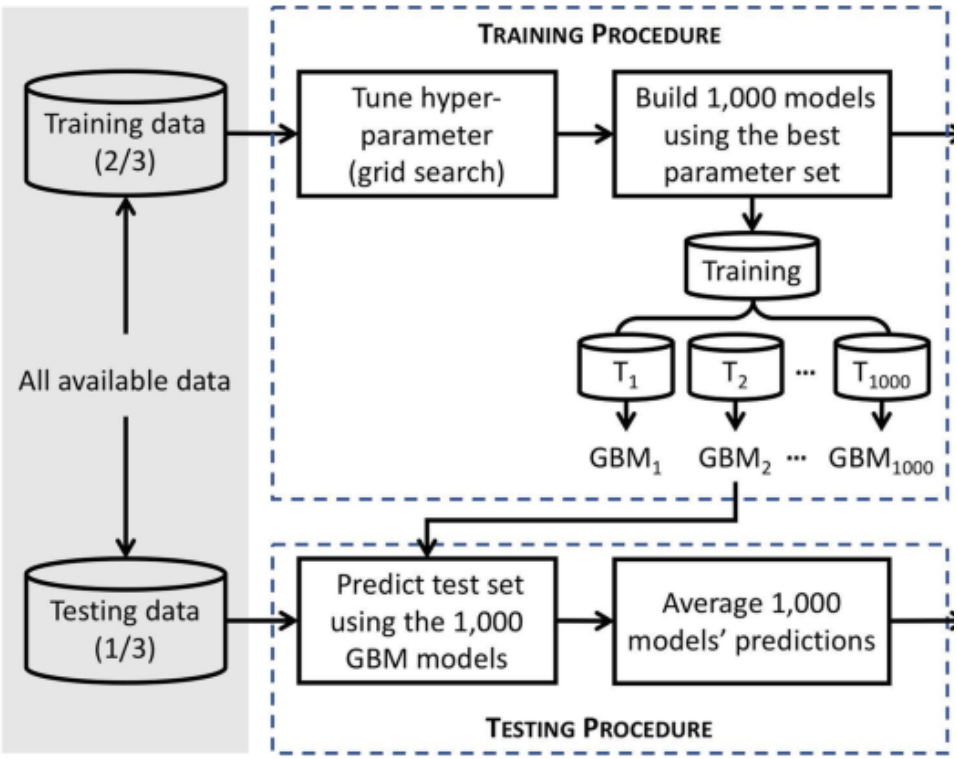

Figure 20: XGBoost Model Architecture *[24]*

4.5 Result Interpretability

Result interpretability is likely one of the most important components in a machine learning-based project. Data science projects are often seen as extremely complex "Black Box" applications capable of producing a prediction for a set of inputs. We address this issue in this project by adding an extensive section on the results' interpretability methods. Furthermore, having a clear understanding of our model's results and features enables us to analyze it for its accuracy and trustworthiness. In general, as a model increases in complexity, it loses interpretability. The same is true if the model uses a large number of features and is subjected to multivariate, non-linear interactions. In real-world applications, interpretable models are more easily accepted as it enables the user to understand model outcomes and override it in cases where it doesn't seem accurate.

Nowadays, there are multiple methods available to add interpretability to a machine learning model. In this project, we have used the following technique:

- Cross correlation plots
- Feature importance chart
- Partial Dependency Plots (PDP)
- P-Values and Coefficients from Ordinary Least Square (OLS)

Cross correlation plots:

Cross correlation plots are excellent way for data interpretation at exploratory data analysis phase. It provides a quick visual reference of the data and highlights some linear interactions between different features. Figure 21 below is an example of one such plot made by using Seaborn library in Python.

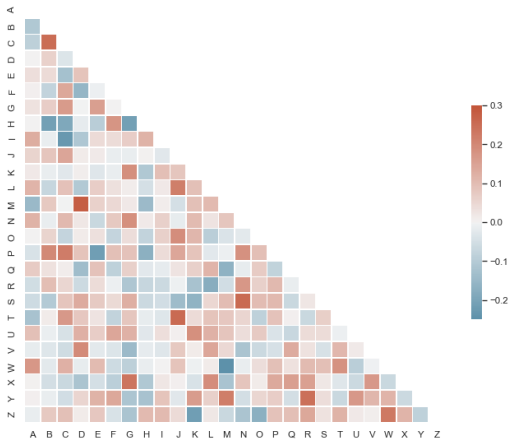

Figure 21: Cross Correlation Heatmap *[25]*

Its relatively straightforward to create such heatmap plots using python libraries like seaborn and matplotlib and they add a great visual data exploration attribute.

Feature importance chart:

The importance of a feature is the increase in the prediction error of the model after we permuted the feature's values, which breaks the relationship between the feature and the true outcome. The concept is straightforward: we measure the importance of a feature by calculating the increase in the model's

prediction error after permuting the feature. A feature is “important” if adjusting its values increases the model error because, in this case, the model relied on the feature for the prediction. A feature is “unimportant” if adjusting its values leaves the model error unchanged because, in this case, the model ignored the feature for the prediction [26].

A feature importance chart is very useful in a machine learning project. It helps a data scientist choose only the features relevant for the model and thus increase overall interpretability. It is very easy to generate using machine learning libraries like Scikit Learn, XGBoost, SHAP etc. Figure 22 below shows an example of a feature importance plot.

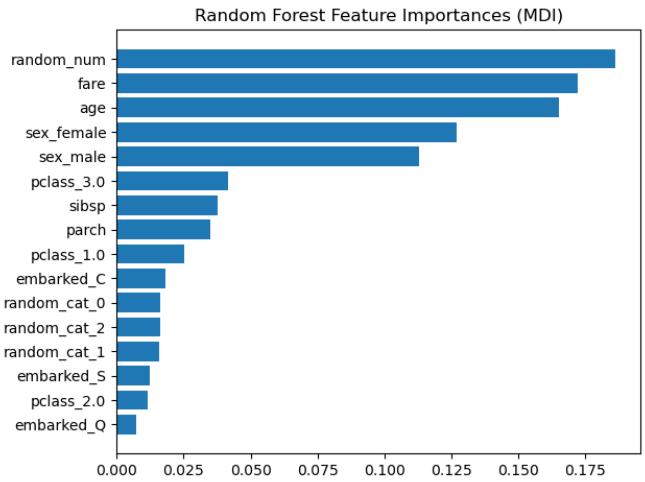

Figure 22: Feature Importance Plot for Random Forest Model *[27]*

Partial Dependency Plots (PDP):

The partial dependence plot (PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model [28]. A PDP is useful in cases where it is important to understand the effect of change in the variable of interest on the overall model outcome. Another feature in the PDP library is the two variable interaction plots. It is similar to the PDP but uses two variables of interest instead of one to interpret the effect of change in a two-variable pair on the final outcome.

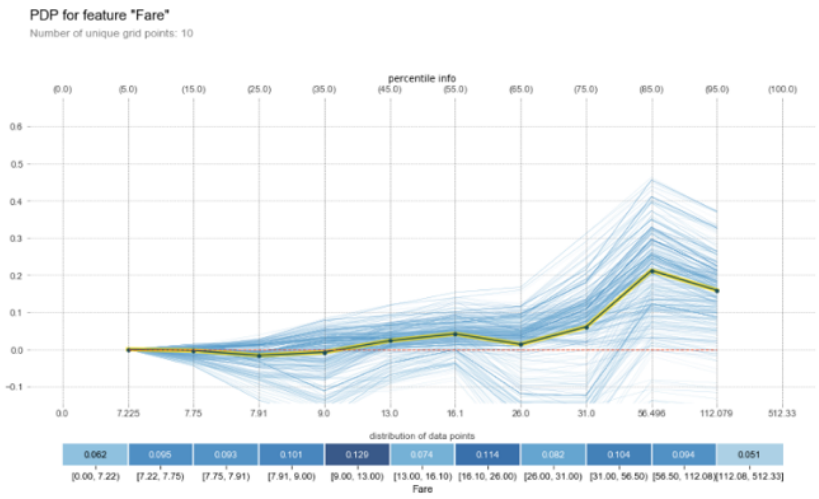

Figure 23: Partial Dependency Plot *[29]*

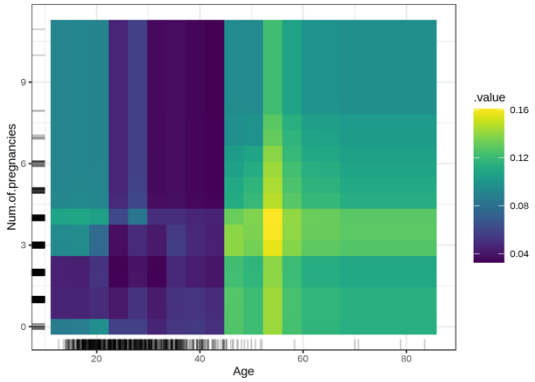

Figure 24: Two Variable Interaction Plot *[26]*

Ordinary Least Square (OLS):

Similar to the feature importance plot, the p-values from an ordinary least square analysis can provide insights into a feature's statistical importance. It is important to note that OLS can only evaluate linear regression relationships. Even with this limitation, in early data exploration phases, OLS p-value and coefficient result can help determine some of our dataset's important features. This is relatively easy to interpret. In general, a p-value < 0.05 suggests that the null hypothesis can be rejected. The null hypothesis in OLS regression is that the feature has no importance on the overall model performance. Thus, a lower p-value signifies higher importance for the feature. The coefficient, on the other hand,

has a slightly different interpretation. These coefficients represent the mean change (or slope) in the outcome for one unit of change in the feature while holding other variables in the model constant.

Chapter 5: Results & Conclusion

In this chapter, we will provide the results and a detailed discussion with a case study. Results presented here are normalized to meet the confidentiality requirements from Amgen. However, the normalization process does not affect the outcomes and learnings in any way.

5.1 Case Study Description

This study was conducted on a mechanical autoinjector that is sold along with a drug prefilled syringe. This product is designed to meet the user needs for the drug that it was dispensing. The model built during this project is flexible with its input features that can expand to various autoinjectors and drug product combinations. The user feedback data was collected for US-only devices and was further filtered to contain feedback specifically related to drug injection time.

The following sections of this chapter will walk through the characteristics of data available, results from the techniques described in the research methodology chapter, discussion on the findings, and then finally present a detailed conclusion.

5.2 First Principle Model Output

In the first principle model, as explained in section 3.2, we are trying to calculate injection time for a given drug product and autoinjector configuration. Equations 3 and 4 from Chapter 3 present the relationship of injection time with design attributes. We used these equations as our guiding principle to formulate a python function that calculated the injection time mean and standard deviation for a given set of design attributes. Below we can see these equations again for reference:

$$F_{plunger} = \left(\frac{2\pi\mu_{oil}\, r_b l_{stopper}}{d_{oil}}\right)\bar{v} + \left(\frac{8\pi\mu l_n r_b^4}{r_n^4}\right)\bar{v}$$

where μ_{oil} is the viscosity of lubricating oil, d_{oil} is the thickness of lubrication layer, $l_{stopper}$ is the length of the stopper in contact with glass, μ is the viscosity of drug product and $\bar{v}$ is the injection speed (linear piston speed with dimensions of length over time).

$$t_{inj} = f(\mu, l_N, r_B, r_N^4, x, 1/(F_{plunger} - F_{friction}))$$

where t_{inj} is injection time, μ is viscosity, l_N is needle length, x is stopper displacement, r_B is inner radius of syringe barrel, r_N is inner radius of needle, $F_{plunger}$ is plunger force, $F_{friction}$ is frictional force and $F_{fluidic}$ is fluid resistance force caused by the pressure drop in the needle.

To validate our model's accuracy, we substituted the design attributes like syringe features and stopper features with actual measured values from our datasets and compared the simulated output with measured experimental injection time values.

It should be noted that although injection time strongly correlates with the needle radius, we could not include this design attribute in the out model as the data was unavailable from the supplier. We assumed the needle diameter to be a constant value for injection time calculation as per the design specification.

We performed a Monte Carlo simulation on the calculated injection time mean and standard deviation to generate simulated injection time values. This technique provided us a normal distribution of 10,000 injection time values. Figure 25 below shows this simulated injection time graph.

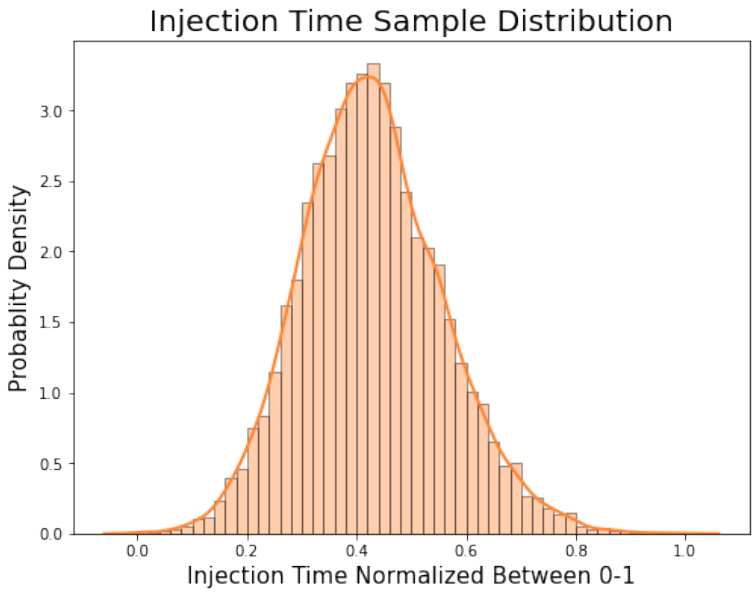

Figure 25: Injection Time Simulated Values

5.3 Exploratory Data Analysis

This project's data came from multiple sources like supplier's database, manufacturing database, user support teams, early product development teams, etc. Broadly the data can be divided into two categories: Features and Target.

Features were data groups used as inputs for the machine learning and first principle models. Target data, in this case, was User Feedback Metric (UFM) which was inversely proportional to the ease of usability of the product. Features were chosen by carefully selecting product design metrics that directly or indirectly affect the product's injection time. We emphasized building our model centered around the injection time feature. Our first principle model was built to predict injection time from design metrics and thus eliminated the need for empirical data for the model's future use case.

Syringe Based Features:

As discussed in section 3.2, pre-filled syringe attributes play an essential role in the drug injection properties. In our analysis, a total of 12 features were used for the syringe design. This feature list

consists of the statistical properties mean, standard deviation, min, and max for three syringe design attributes.

Stopper Based Features:

The stopper is a critical component of the autoinjector assembly. It acts as a barrier between the drug product and the syringe plunger. Like the syringe, we used 12 stopper features, consisting of the four statistical properties of 3 design attributes. Figure 26 shows the syringe and stopper positions in an autoinjector.

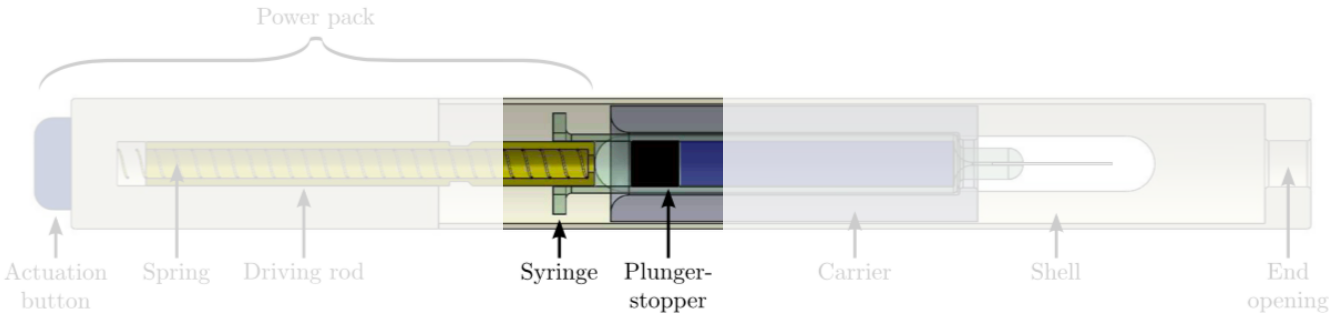

Figure 26: Syringe and Stopper in an Autoinjector

Drug product features such as viscosity and density determine the limit of possible injection time. Highly viscous drugs require more force from the power pack subunit to overcome the opposing frictional and fluid resistance forces. In our project, we have used four drug product related features. The statistical properties of drug product features were not considered as this is not directly associated with a product design feature and thus doesn't have a typical engineering limit.

Spring Based Features:

The spring in an autoinjector is a crucial component in the power pack. A power pack is essentially the subunit that powers the autoinjector and drives the drug from the syringe to the patient through a

needle. The spring stores the mechanical energy needed to drive this. Springs are thus characterized based on the force they can produce due to the stored energy.

For this analysis, we used a total of 10 spring or power pack-based features. In addition to the spring forces, we also included features related to the functional parameters for the power pack.

Functional Parameters Based Features:

Unlike the features described above, functional parameter-based features are only used in the User Feedback Metric (UFM) predictive modeling. These features are the functional output of an autoinjector and hence characterized as functional parameters. These features are used to describe an autoinjector's performance. Injection time, a functional parameter, acts as a bridge between the design features and the UFM. Figure 26 demonstrates this relationship graphically.

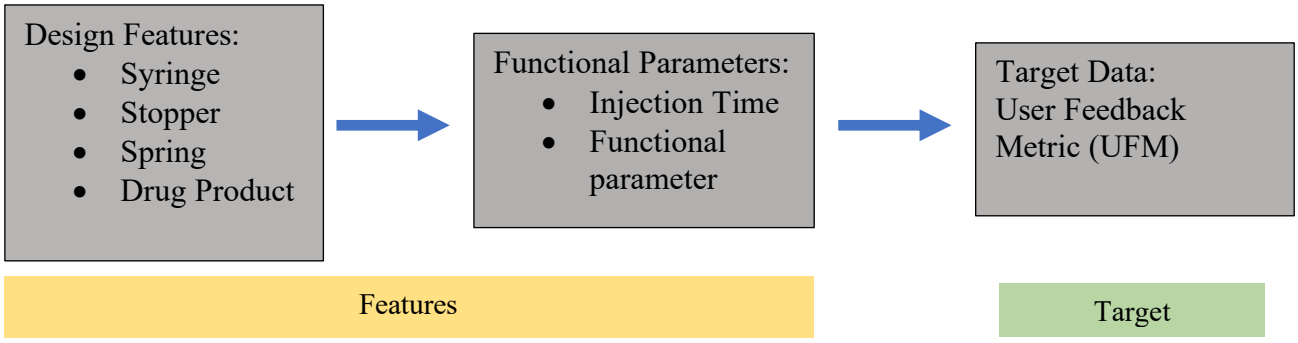

Figure 27: Data Categories & Relationships

In addition to injection time, we have included one more functional parameter in the form of its statistical properties. Thus, we have used a total of 8 features based on the functional parameters.

Correlation Relationships:

As the first step in exploratory data analysis, we created correlation plots between the features and the target variables to get intuitive relationship patterns. Figures 28 – 32 below shows a correlation plot of different categories of features. We chose to plot these graphs separately to make them easier to read.

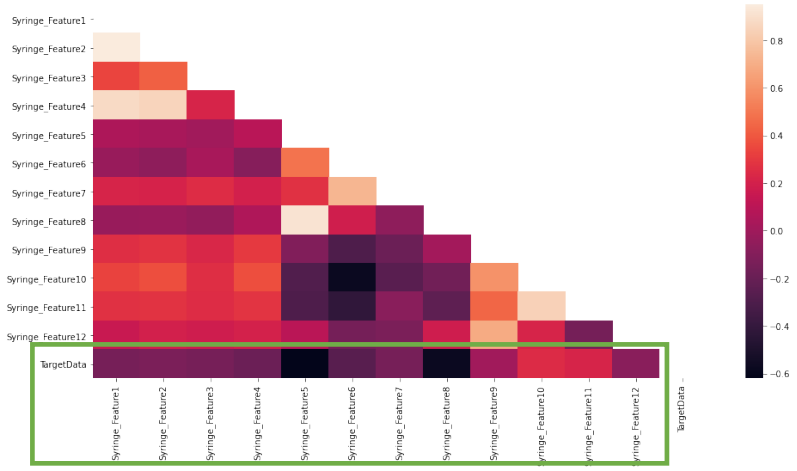

Figure 28: Syringe Features Correlation Plot

We see that most of the parameters here have no direct correlation with our target data in the syringe feature correlation plot. Some features like the Syringe Feature 5 and Syringe Feature 8 have a moderately strong negative correlation with the target data, i.e., with an increase in these parameters, we see a decrease in the target data UFM. As explained earlier, the lower the UFM, the better the product performance is. Similarly, we see that Syringe Feature 10 and 11 have a moderate positive correlation with target data.

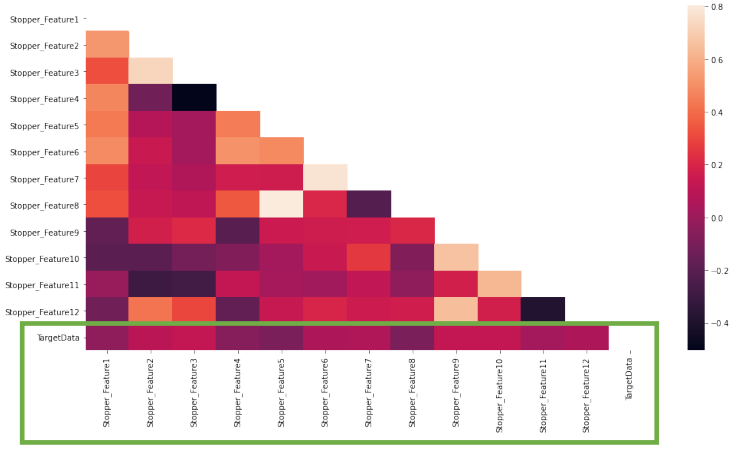

Figure 29: Stopper Features Correlation Plot

Unlike the syringe features, the stopper features don't have a negatively correlated variable. This means that any increase in the feature values will increase the UFM value and decrease overall user experience even though the product is performing within technical specifications.

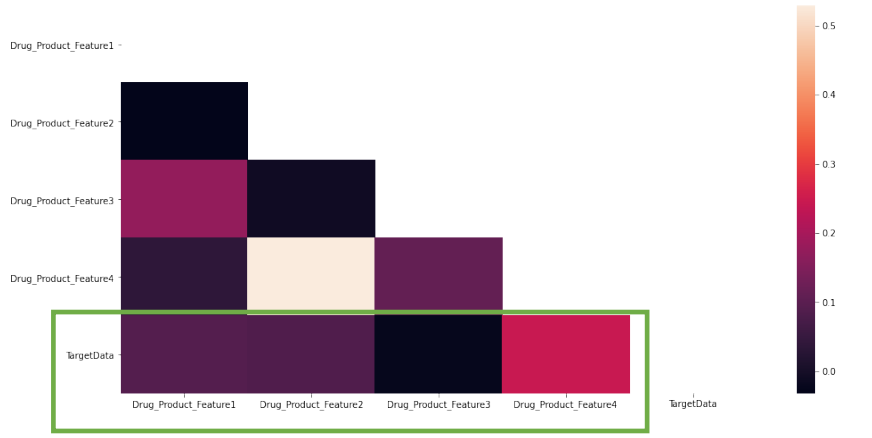

Figure 30: Drug Product Features Correlation Plot

For drug product cased features, we see a very straightforward relationship. Out of the four variables, we see that only Drug Product Feature 4 has a noticeable positive correlation with target data.

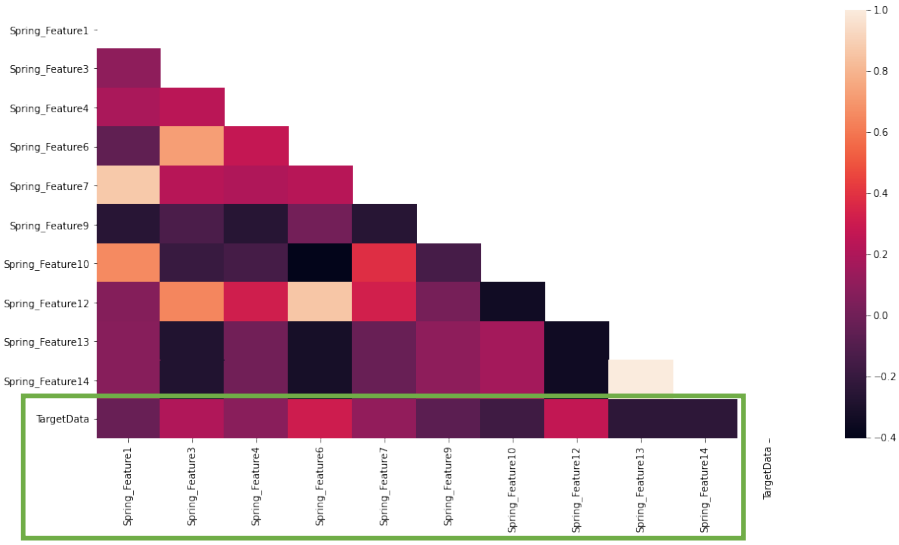

Figure 31: Spring Features Correlation Plot

For spring-based features, we see that Spring Feature 13 and 14, a representation of spring force, is negatively correlated with our target data. This makes sense conceptually, as an increase in the spring force relates to a faster injection time which enhances user experience.

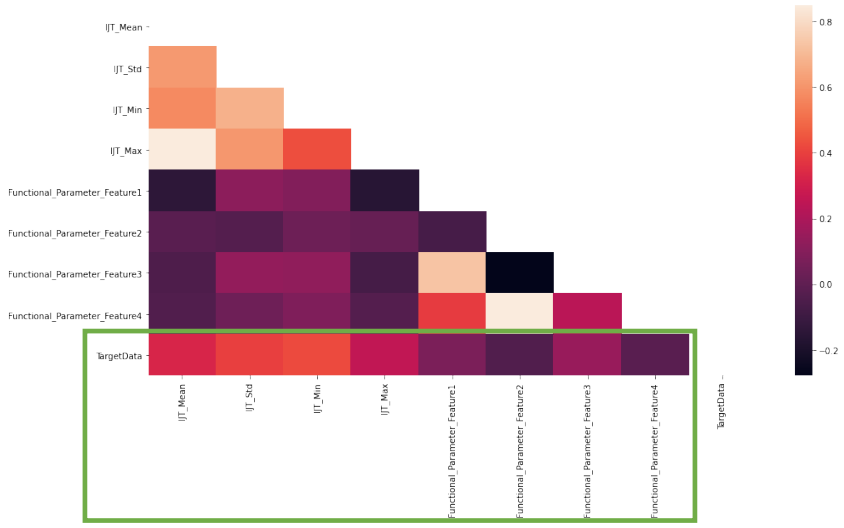

Figure 32: Functional Parameter Features Correlation Plot

For Functional Parameters based features, we see that the Injection Time is moderately strongly correlated with the target data. The strongest among the Injection time features is the Injection_Time_Min which is the minimum value of injection time measured out of the samples in a given batch. This also makes sense intuitively as the variations in injection time minimum usually mean that although these values were within the product specification limits, longer injection times can often be perceived for user discomfort. Another scenario with this variable could be a device failure or stalling, resulting in a higher minimum injection time and resulting in a negative user experience.

In conclusion, with the correlation analysis, we were able to identify some critical variables in our dataset that can be useful in understanding our model and its relationship with the variables.

Ordinary Least Square (OLS) Relationships:

Ordinary least square is a fundamental analysis method useful in exploring linear relationships between the predicting features and target variables. In our case, since the variables have non-linear dependencies, we can't use the model results directly. However, the OLS regression result helps us in

understanding variable importance by checking its coefficient value. Table 1 below gives a brief output (coefficients) from the OLS regression method.

Table 1: OLS Regression Result

Feature Name	Coeff	Feature Name	Coeff	Feature Name	Coeff
Syringe_Feature1	-0.3719	Stopper_Feature1	0.1636	Spring_Feature7	0.1323
Syringe_Feature2	0.3677	Stopper_Feature2	0.029	Spring_Feature9	0.0029
Syringe_Feature3	-0.1478	Stopper_Feature3	-0.1858	Spring_Feature10	-0.0181
Syringe_Feature4	-0.0859	Stopper_Feature4	-0.106	Spring_Feature12	0.077
Syringe_Feature5	0.181	Stopper_Feature5	-0.1658	Spring_Feature13	-5.71E+07
Syringe_Feature6	-0.0408	Stopper_Feature6	0.132	Spring_Feature14	5.71E+07
Syringe_Feature7	-0.1852	Stopper_Feature7	-0.0548	IJT_Mean	0.0551
Syringe_Feature8	-0.2679	Stopper_Feature8	0.0236	IJT_Std	0.0015
Syringe_Feature9	0.1759	Stopper_Feature9	0.1656	IJT_Min	0.006
Syringe_Feature10	-0.1072	Stopper_Feature10	0.0753	IJT_Max	-0.016
Syringe_Feature11	0.0527	Stopper_Feature11	-0.1649	Functional_Parameter_Feature1	-0.0931
Syringe_Feature12	-0.0089	Stopper_Feature12	-0.1824	Functional_Parameter_Feature2	2.0231
Drug_Product_Feature1	0.1172	Spring_Feature1	-0.1945	Functional_Parameter_Feature3	0.8801
Drug_Product_Feature2	-0.1115	Spring_Feature3	-0.2227	Functional_Parameter_Feature4	-1.9999
Drug_Product_Feature3	-0.0031	Spring_Feature4	0.0169		
Drug_Product_Feature4	0.1446	Spring_Feature6	0.1563		

In Table 1, we can see a similar trend in the variable importance; however, it is not perfect, which confirms our hypothesis that the variables have a non-linear relationship with the target variable. Thus, in conclusion, we further improved our model to include non-linearity using decision tree types of models. Using a tree-based model as opposed to a convoluted neural network or deep learning model was to improve the model accuracy and preserve the interpretability.

5.4 Random Forest Regressor

Our first decision tree-based model was a Random Forest model. We used the Scikit Learn Random Forest Regressor package to perform these analyses. Before the analysis, we prepared the data set by normalizing all the features and target data set between 0 and 1. Finally, we divided the entire data set into training and testing data sets and performed a 5-fold cross-validation. Our entire data set consisted of total of 270 unique batches. Each batch consists of multiple units, and the design and manufacturing based variables were measured and aggregated by batches. We used 75% of this dataset for training and the remaining for testing the model accuracy.

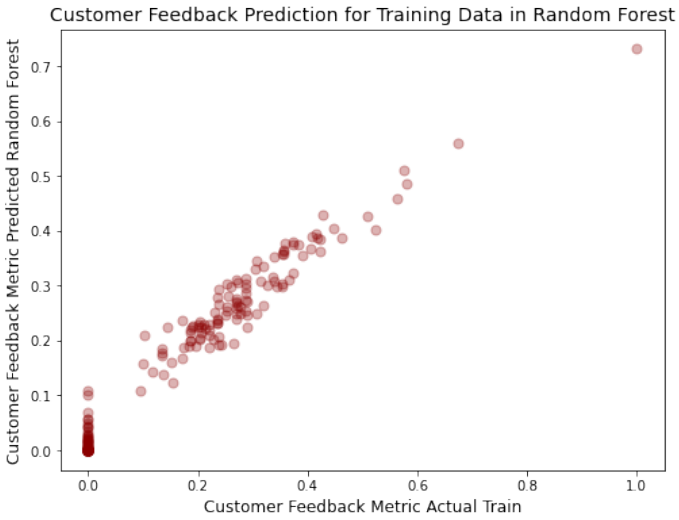

Figure 33: Random Forest Model Output for Training Data

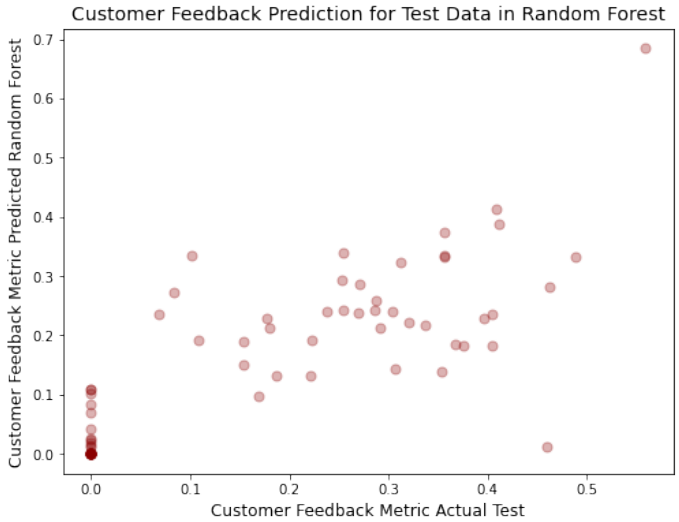

Figure 34: Random Forest Model Output for Test Data

As observed from the data, our training set has a very high R-square value (0.95) which is the metric for model accuracy. This is not surprising as the model was built and trained on this data set. However, when we see the test data set, we observe some variation in the predicted data versus the actual data. The R-square for the test data set with Random Forest was 0.6. Figure 35 below shows a visualization of the random forest estimator trees for our model.

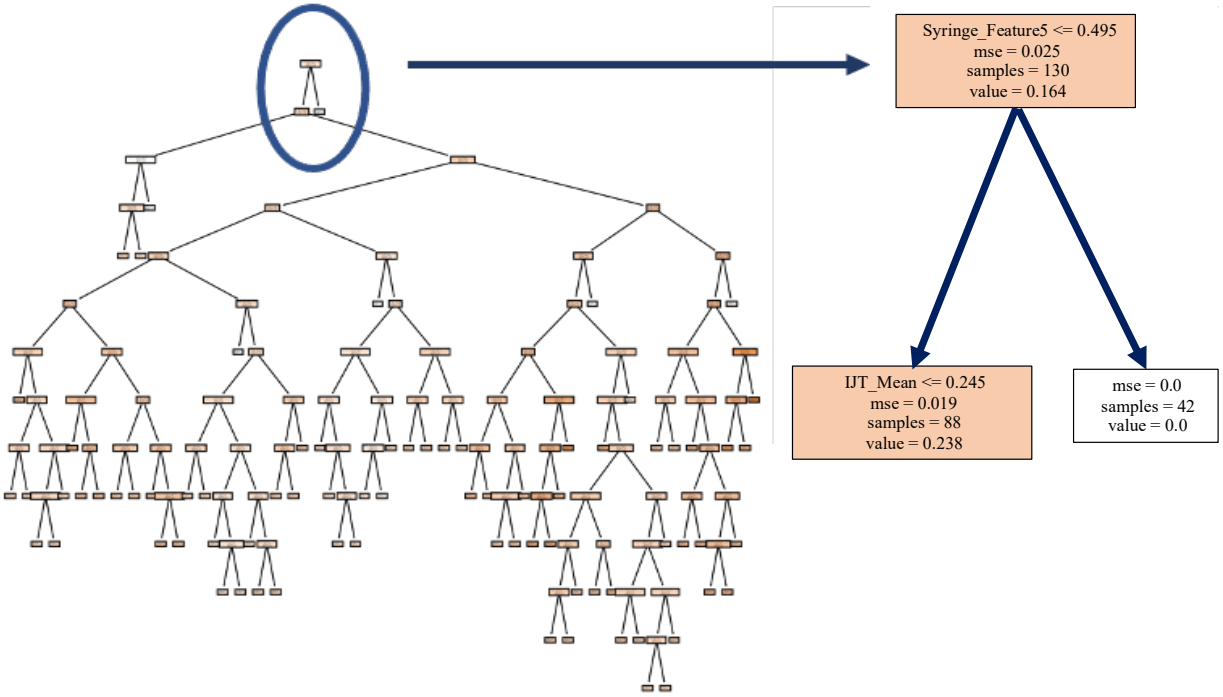

Figure 35:: Random Forest Estimator Tree (Left) and an Enlarged Image of one of its leaf (Right)

Although the Random forest gave us decent accuracy for the model, we wanted to increase the model complexity further to see if we could gain more accuracy.

5.5 Gradient Boost Regressor

Our next tree-based model was Gradient Boost Model. Gradient Boosting trains many models in a gradual, additive, and sequential manner. We used a similar approach for our data as mentioned above for Random Forest, with 75% of the data set used for training and the remaining for testing. In addition to fitting the training data set into this algorithm structure, we also performed hyperparameter optimization. We used grid search and a 5-fold cross-validation technique to optimize the hyperparameters. The selection criteria for the best model were based on the lowest mean squared error. Figure 36-37 below shows the performance of a Gradient Boost Model. As it can be easily seen

that the training data set is highly correlated here with R-Square of 0.91, similar to what we saw in the case of Random Forest and again, the R-square for the test data set was 0.59. It is evident that even with increased complexity, we have not gained any significant improvement in the model accuracy; on the contrary, it has become worse.

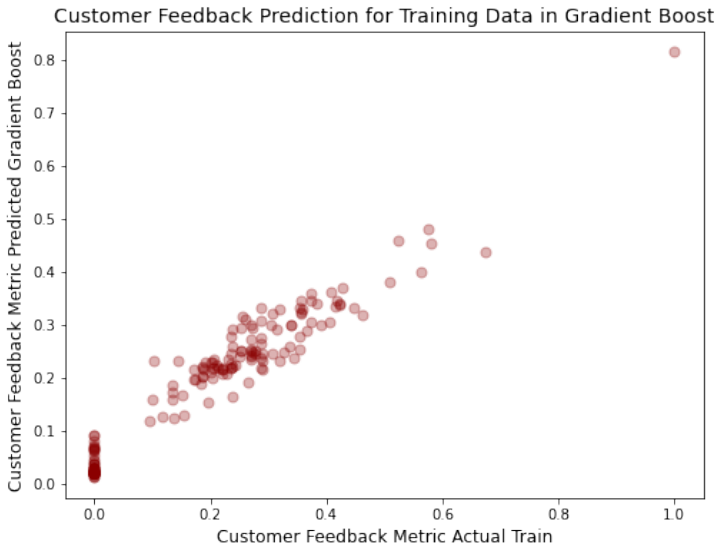

Figure 36: Gradient Boost Model Output for Training Data

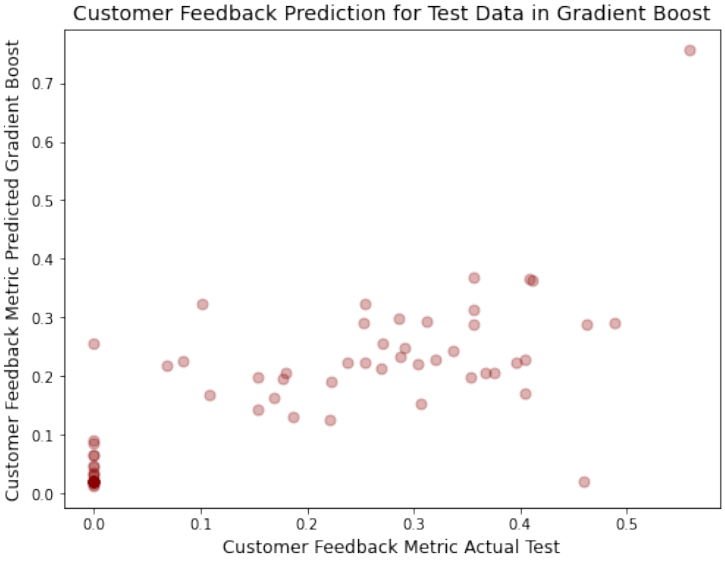

Figure 37: Gradient Boost Model Output for Test Data

5.6 XGBoost Regressor

The third tree-based model that we used was the XGBoost regressor. Theoretically, this is an extension of the Gradient Boosting Model, which is now optimized to the computational power by performing a parallel tree formation process. Unlike Gradient Boosting, XGBoost provides a more holistic bias and variance trade of balance by focusing on regularization factors. Thus, XGBoost surpasses Gradient Boosting for practical implementation to have a stable and robust machine learning framework.

For this project, we performed a feature importance analysis with XGBoost. Figure 38 below shows a chart of the top 10 features ranked in the order of their importance in the model performance.

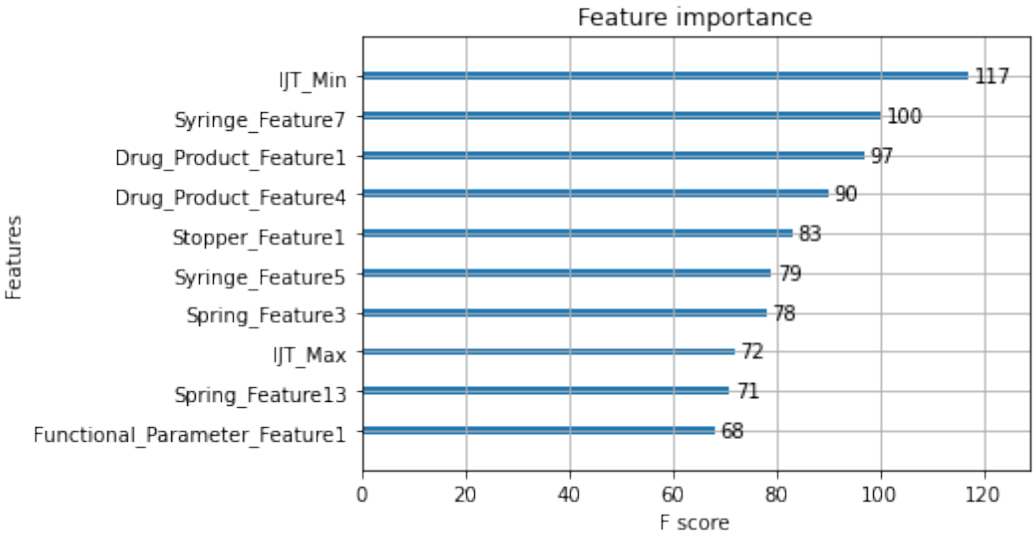

Figure 38: Feature Importance Chart – XGBoost (IJT: Injection Time)

With XGBoost, we used the same data set with 75% for model training and 25% for testing. We optimized the hyperparameters for XGBoost by performing a Bayesian Optimization method. Bayesian Optimization works by building a probabilistic model of the objective function (surrogate function) and then efficiently searching it with an acquisition function before candidate samples are chosen to evaluate the actual objective function. Figure 39-40 below shows the prediction plot of training and test data with the XGBoost model. The model performance was close to the other models with R-Square for training set as 0.75 and testing set as 0.56.

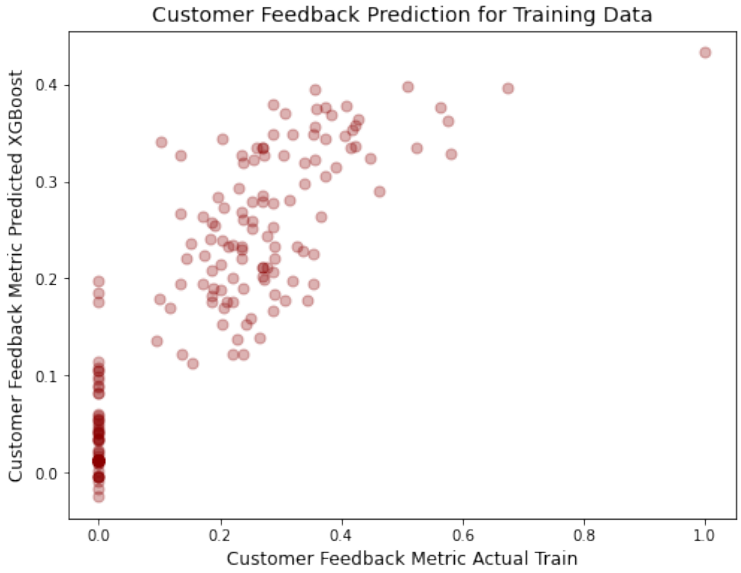

Figure 39: XGBoost Prediction Plots on Training Data

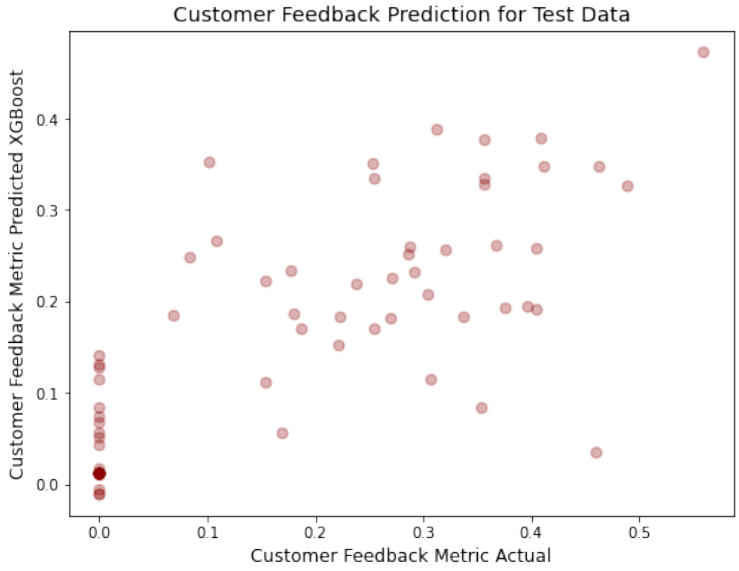

Figure 40: XGBoost Prediction Plots on Training Data

It's interesting to note that unlike the Random Forest or Gradient Boost model, XGBoost has a much lower training set R-square (0.75). This makes intuitive sense as, by its definition, the XGBoost model tries to penalize for overfitting, creating a much better balance between the bias and variance trade-offs. Thus, XGBoost is the best performing model for applications where model reliability is critical.

5.7 Results Interpretability using Partial Dependency Plots

One of the key objectives of this project was to directly relate the impact of change in injection time features on the target variable. With all the results presented thus far, we can confirm that the model built using decision tree-based algorithms provides reliable and moderately accurate predictions for our data. The model was constructed using more than 40 features from multiple design attributes related to the Syringe, Stopper, Drug Product, Spring, and the device's Functional Parameters. However, it is essential for design teams to predict the impact of changes in injection time alone on the target data. Injection time is a functional parameter that is observed directly by the user. It is not the only functional parameter that is important for the design teams at Amgen. However, in this project, injection time was the critical feature that can be calculated with first-principle models and became a bridge between the design specifications and user feedback. To fulfill this task, we implemented the partial dependence method on our machine learning models. Figure 41 below shows a partial dependence plot for the feature Injection Time mean. In the figure below, the yellow line denotes the average effect of a feature. Simultaneously, the individual green lines known as the individual conditional expectation (ICE) plots display one line per instance that shows how the instance's prediction changes when a feature changes.

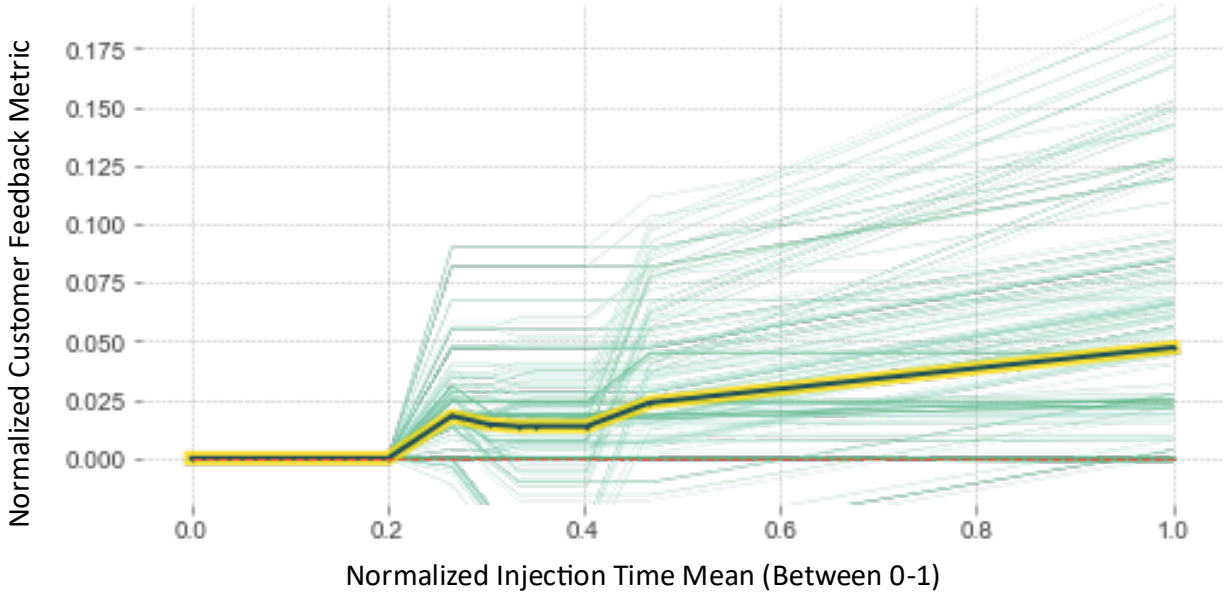

Figure 41: Partial Dependency Plot of Normalized Injection Time Mean Value (Between 0-1)

Next, we created the partial dependence plot for Injection Time minimum as it was identified as one of the important features for our model. Figure 42 below shows the plot.

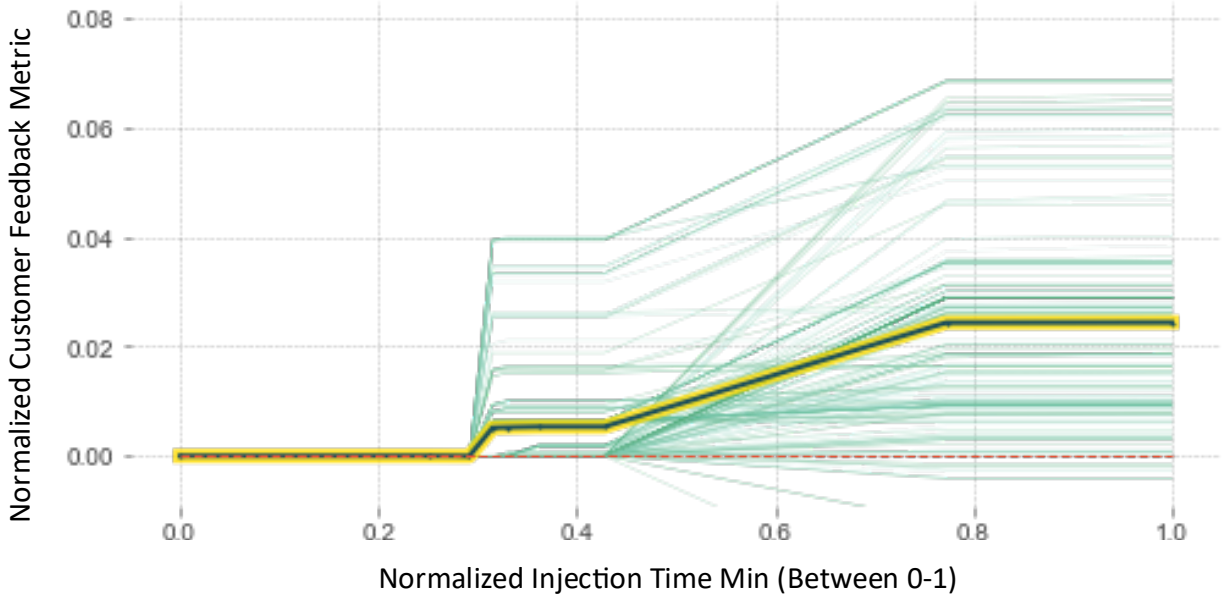

Figure 42: Partial Dependency Plot of Normalized Injection Time Minimum Value (Between 0-1)

After comparing the two graphs above, we can easily see that the Injection Time mean variation has a much larger impact on the user feedback metric than the Injection Time minimum feature. In addition to analyzing the individual features as partial dependency plots (PDP), we also created a PDP interaction plot that can plot 2 features as a 2-dimensional heatmap. In this heatmap, each data point illustrates a combination of two variables and their effect on the target variable. Values in the chart represent the user feedback metric. As discussed earlier, higher value of user feedback metric represents negative user experience. Figure 43 below shows the PDP interaction plot of Injection Time Mean with Injection Time Minimum.

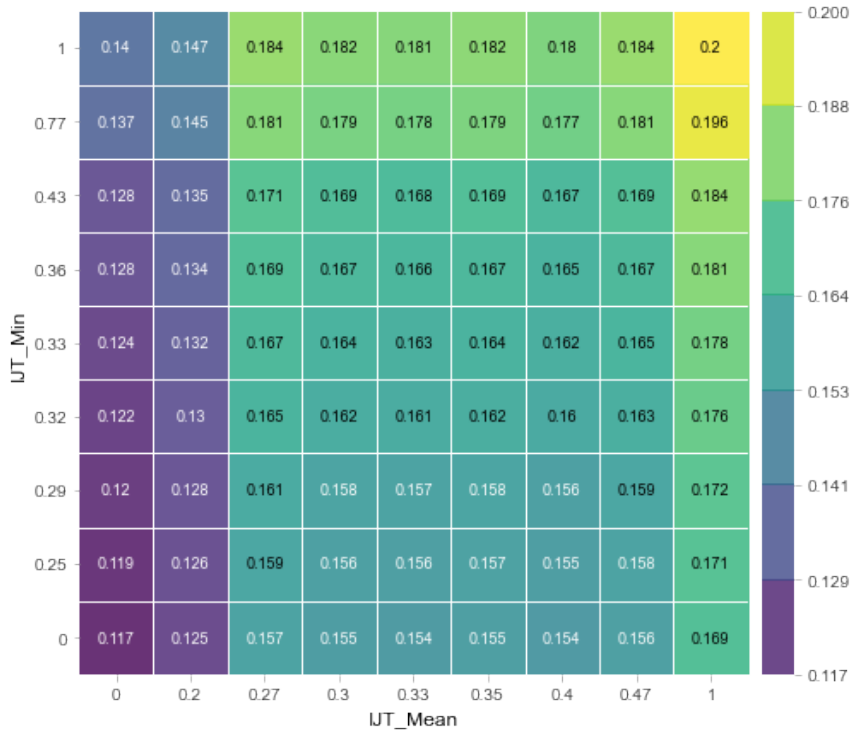

Figure 43: Partial Dependency Interaction Plot of Injection Time Mean and Injection Time Minimum

5.8 Conclusion

In this project, we started with an objective to create a link between the data generated at the product design and development phase to the end user feedback metric. This type of correlation was never

done before to the best of our knowledge. We began the analysis by converting the design specification metrics into a device functional parameter with the first principle modeling. This also helped us create synthetic data to be used in the machine learning models. From this first phase analysis, we learned that it is possible to create a reliable and robust relationship between design specifications and the device's injection time. Our results show that the variation between the simulated and actual injection time was approximately 12%. This is a good verification of our first principle model, considering that we don't account for manufacturing and assembly variations in our simulation model. After this step, we began our machine learning model. To build a robust and accurate model, we chose over 40 parameters related to key design and functional attributes and their statistical components like mean, standard deviation, minimum and maximum values. We decided to build and train our machine learning model on empirical data to incorporate the inherent noise or variations during the manufacturing and assembly process. After comparing several decision tree-based models, we concluded that the XGBoost model provided us with the best balance between model accuracy and robustness. The other two models were more biased and appeared to be overfitting in the training data set. Our final phase of analysis was to build an easy to interpret feature that can help design teams use our anchor functional parameter i.e., injection time, and precisely measure the variations in the user feedback metric directly caused by changes in injection time parameter. We performed a partial dependence analysis on our machine learning model and created partial dependence and interaction plots to achieve this. These plots provided an easy-to-read metric for design teams to quickly understand the impact that a change in design feature could have on user experiences.

In conclusion, we can confirm that the project was successful in achieving the set objectives. Apart from the technical benefits, this project also delivers substantial business benefits by providing an

early validation of user experience without an actual device that ultimately saves time and resources required.

5.9 Learnings and Recommendations for Future Work

As a recommendation for future work, we would like to suggest the following considerations to improve the model performance.

Data Availability:

Although we started with a large set of user feedback data, after collating it with other manufacturing and supplier data for design attributes, we lost a significant amount as the data common between these databases was relatively sparse. For future work, we would recommend to work more closely with the suppliers and integrate data at a much regular cadence.

Experimental Consistency Between Early Development and Manufacturing Data:

We had a significantly rich data pool available from the early development phases where each of the design components was thoroughly investigated for its impact on the overall device performance. Although this helped get a theoretical understanding of the device and its component's significance, we could not use it directly to train our machine learning models as these characterization techniques were not the same as the ones used in manufacturing lines. We understand that this may be driven by the early development and the production teams having less visibility on their data acquisition and experimental set up techniques.

Enhanced User Feedback Data Labeling:

Finally, we had a great collection of data from the post-market surveillance activities in the form of user feedback metric. This data was most abundant as Amgen invests many resources in providing the best experiences for their users. However, there were some opportunities where this database can be further enhanced, especially if it is being used for a machine learning or data science project. Currently,

this database consists of a lot of information in a format that is ideal for manual data retrieval and analysis. To make it easier for the machine to process, we need to add few additional fields to help with a more straightforward classification. One way to implement such a technique would be to classify the data based on hierarchical yes-no based questions into multiple originating categories.

References

[1] R. A. Rader, "(Re)defining biopharmaceutical," *Nature Biotechnology,* pp. 743-751, 2008.

[2] Y.-C. Chen and M.-K. Yeh, Introductory Chapter: Biopharmaceuticals, IntechOpen, 2018.

[3] C. C. Quianzon and I. Cheikh, "History of Insulin," *Journal of Community Hospital Internal Medicine Perspectives ,* pp. 1-3, 2012.

[4] P. Agrawal, "Biopharmaceuticals: An emerging trend in Drug Development," *SOJ Pharmacy & Pharmaceutical Sciences,* pp. 1-2, 2015.

[5] G. Jagschies, "Brief Review of the Biopharmaceutical and Vaccine Industry," in *Biopharmaceutical Processing*, Elsevier Ltd., 2018, pp. 33-58.

[6] A. Batta, B. S. Kalra and R. Khirasaria, "Trends in FDA drug approvals over last 2 decades: An observational study," *J Family Med Prim Care,* pp. 105-114, 2020.

[7] T. Marrow and L. H. Felcone, "Defining the difference: What Makes Biologics Unique," *Biotechnology Healthcare,* vol. 1, no. 4, pp. 24-29, 2004.

[8] Research and Markets, "The Global Biopharmaceuticals Market is Projected to Grow by 8.7% Through to 2025 and Reach $446 Billion - ResearchAndMarkets.com," June 2019. [Online]. Available: https://www.businesswire.com/news/home/20190624005782/en/Global-%20Biopharmaceuticals-Market-Projected-Grow-8.7-2025.

[9] Amgen Inc, "Amgen History," 2021. [Online]. Available: https://www.amgen.com/about/amgen-history.

[10] Amgen Inc, "About Amgen," Amgen, 2020.

[11] S. Neadle, "Combination Product Development Challenges," International Pharmaceutical Quality: Inside the Global Regulatory Dialogue.

[12] K. Sturgis, "How Autoinjector Technologies Could Change Drug Delivery. Retrieved from Medical Device and Diagnostic Industry," June 2018. [Online]. Available: https://www.mddionline.com/design/how-autoinjector-technologies-could-change-drug-delivery.

[13] Manufacturing Chemist, "Autoinjector Development," 7 March 2012. [Online]. Available: https://www.manufacturingchemist.com/news/article_page/Autoinjector_developmen%20t/76465.

[14] P. J.Ogrodnik, "Postmarket Surveillance," in *Medical Device Design*, Elsevier Ltd., 2013, pp. 287-298.

[15] FDA CDRH, "Design Control Guidance for Medical device Manufacturers," FDA, 1997.

[16] "FDA Approval Pathway for Medical Devices - Part 2," [Online]. Available: https://www.imarcresearch.com/blog/part-two-fda-approval-pathway-for-medical-devices.

[17] J.-C. Velleux, "Pressure and Stress Transients in Autoinjector Devices," CALIFORNIA INSTITUTE OF TECHNOLOGY , Pasadena, California, 2019.

[18] P. K. Kundu, I. M. Cohen and D. R. Dowling, Fluid Mechanics. 5th ed., Oxford U.K: Elsevier, 2012.

[19] B. L. R. D. Thueer T, "Development of an advanced injection time model for an autoinjector.," *Med Devices (Auckl),* vol. 11, pp. 215-224, 2018.

[20] P. P. B. E. e. a. Nitin Rathore, "Variability in Syringe Components and its Impact on Functionality of Delivery Systems," *PDA J Pharm Sci and Tech,* vol. 65, pp. 468-480, 2011.

[21] A.-B. S. et.al., *JMR Med Inform,* vol. 6, no. 2, p. 34, 2018.

[22] Databricks, "Databricks Documentation : Notebooks," [Online]. Available: https://docs.databricks.com/notebooks/index.html.

[23] H. et.al., "Should You Derive? Or let the data drive? Towards a First Principles Data Driven Symbiosis," in *IMA*, MN, 2016.

[24] D. M. U. A. B. Q.-J. L. Y. Z. L. L. Yuanyuan Li, "Putative biomarkers for predicting tumor sample purity based on gene expression data," *BMC Genomics,* vol. 20, no. 1021, 2019.

[25] Seaborn, "Seaborn Examples," [Online]. Available: seaborn.pydata.org.

[26] C. Molnar, "Permutation Feature Importance," in *Interpretable Machine Learning*, bookdown, 2021.

[27] Scikit Learn, "Permutation Importance vs Random Forest Feature Importance (MDI)," [Online]. Available: https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-py.

[28] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Annals of Statistics ,* vol. 29, no. 5, pp. 1189-1232, 2001.

[29] X. B. Sicotte, "How to create partial dependency plot for logistic regression in Python sklearn," 11 2018. [Online]. Available: https://stats.stackexchange.com/q/376787.

[30] F. Scherer, "R&D Costs and Productivity in Biopharmaceuticals," Mossavar-Rahmani Center for Business & Government, Harvard Kennedy School, 2011.