Leveraging Predictive Analytics to Assess Operations Metrics

by

Nicholas Charles Samuel Artman
BS in Supply Chain Management, The Pennsylvania State University
and
Chi-Wei Kong
BS in Supply Chain Management, Finance, Information Management & Technology, Syracuse University

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2021

Signature of Author: _____
Department of Supply Chain Management
May 14, 2021

Signature of Author: _____
Department of Supply Chain Management
May 14, 2021

Certified by: _____
Maria Jesús Saenz, PhD
Executive Director, Supply Chain Management Blended Program
Capstone Advisor

Certified by: _____
Ozden Tozanli Yilmaz, PhD
Postdoctoral Associate
Capstone Co-Advisor

Accepted by: _____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Leveraging Predictive Analytics to Assess Operations Metrics

by

Nicholas Charles Samuel Artman

and

Chi-Wei Kong

Submitted to the Program in Supply Chain Management
on May 14, 2021 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

Many firms rely on key performance indicators (KPIs) to manage their business. Though countless metrics exist, it can be difficult for companies to identify which metrics are driving their performance. This is problematic within industry, as insignificant KPIs can lead to misguided management insights. This research analyzes how companies and organizations can assess operational metrics utilizing predictive analytics. Additionally, it shows how firms can leverage their metrics to prepare for future objectives and identify key predictive indicators. This analysis comprises an assortment of predictive modeling techniques to evaluate manufacturing and inventory metrics for a firm identified as *Company XYZ*. These modeling approaches include multiple linear regression, random forest, LASSO regression, and backward elimination. Our analysis found four of the ten performance metrics reviewed to be significant in predicting a production efficiency metric. Using these four metrics, we applied a multiple linear regression model to assign coefficients that could be leveraged for sensitivity analysis. Our results ultimately identified key predictive indicators and created sensitivity analysis to help management teams prepare for future endeavors. This research demonstrates that predictive analytics can be used as a fast and cost-effective approach for companies to review their performance metrics.

Capstone Advisor: Maria Jesús Saenz, PhD
Title: Executive Director, Supply Chain Management Blended Program

Capstone or Thesis Co-Advisor: Ozden Tozanli Yilmaz, PhD
Title: Postdoctoral Associate

ACKNOWLEDGMENTS

We would like to thank our capstone advisors, Dr. Maria Jesús Saenz, Dr. Ozden Tozanli Yilmaz, and formerly Dr. Nima Kazemi for their continued support throughout our research. This capstone was greatly enhanced by their extensive knowledge in the fields of machine learning, supply chain digital transformation, and manufacturing excellence. We will be forever grateful for their diligence and commitment to our educational experience at MIT.

TABLE OF CONTENTS

1. INTRODUCTION 1

1.1. Problem Statement 2

2. LITERATURE REVIEW 3

2.1. Key Performance Indicators in Supply Chain Management 3

2.2. Current Methods for KPI Review and Selection Utilizing 5

2.2.1. Multi-Criteria Decision Making 5

2.2.2. Linear Programming 6

2.3. Machine Learning Models for Feature Selection and Ranking 7

2.3.1. Multiple Linear Regression 8

2.3.2. Random Forest 8

2.3.3. Backward Elimination 9

2.3.4. Least Angle Selection and Shrinkage Operator (LASSO) 10

2.4. Literature Review Summary 10

3. DATA AND METHODOLOGY 10

3.1. Exploratory Data Analysis 11

3.1.1. Distribution of Variables 12

3.1.2. Correlation of Variable 13

3.1.3. Removing Outliers 14

3.1.4. Scaling Variables 15

3.2. Modeling 16

3.2.1. Backward Elimination 16

3.2.2. LASSO 17

3.2.3. Multiple Linear Regression 18

3.2.4. Random Forest 19

3.3. Sensitivity Analysis 20

3.4. Methodology Summary 21

4. RESULTS 22

4.1. Comparing the Models 22

4.1.1 Cross-Validation 24

4.2. Addressing Multicollinearity 25

4.2.1. Factor Analysis 27

4.3. KPI Selection 31
4.3.1. Addressing Insignificant Variables 32
4.3.2. Interaction Effect 33
4.4. Sensitivity Analysis 33
4.5. Results Summary 35
5. DISCUSSION 36
5.1. Managerial Insights 36
5.2. Furthering the Analysis 38
5.2.1. Replicating the Analysis 39
5.2.2. Potential Improvements 40
5.3. Discussion Summary 42
6. CONCLUSION 42
REFERENCES 45

LIST OF FIGURES

Figure 1: Variable Histograms 12
Figure 2: Correlation Matrix 14
Figure 3: mtry Optimization 20
Figure 4: Random Forest (Preliminary Results) 20
Figure 5: Preliminary Results Summary 23
Figure 6: Cross Validation MAE Comparison 24
Figure 7: Cross Validated Results of Models When Excluding One of the Correlated Variables 26
Figure 8: Factor Loading (Inventory) 28
Figure 9: Factor Loading (Manufacturing) 28
Figure 10: Cross Validated Results of Models Including Inventory Factor 30
Figure 11: Final Model Variable Importance 32
Figure 12: Impact of Output Variable with Changing Single Individual Variable 34

LIST OF TABLES

Table 1: Backward Elimination (Preliminary Results) 17
Table 2: LASSO (Preliminary Results) 18
Table 3: Multiple Linear Regression (Preliminary Results) 19
Table 4: MLR Variable Importance 19
Table 5: Summary of Preliminary Results 23
Table 6: Model Results Excluding Variable I_3 25
Table 7: Model Results Excluding Variable I_1 25
Table 8: Manufacturing & Inventory Factors 29
Table 9: Model Results Using Inventory Factor 30
Table 10: Final Model Results 31
Table 11: Interaction Effect Results 33

1. INTRODUCTION

In the current global economy, some of the most powerful and profitable firms such as Amazon and Facebook have leveraged their analytical capabilities to drive competitive advantage through customer insights and improved operations (Henke et al., 2016). With these firms showcasing the power of these new technologies, other businesses are investing time and assets in developing their capabilities within analytics, big data, machine learning, and many others.

One field of study in which these new technologies can be applied is supply chain management. Through increased accuracy of demand forecasting using machine learning and improved warehousing practices leveraging real-time analytics, supply chain management has used many of these capabilities to increase operational efficiency across industries (Flovik, 2019). Although these technologies help drive value, they can also increase the complexity of a firm's supply chain. For this reason, many business leaders seek simple and measurable metrics when managing their teams in a highly digitized world. Key performance indicators allow businesses to quantify and assess the health of the different functions within their business. Understanding the need for clarity and the importance of performance metrics, supply chain professionals are exploring options to use machine learning to best-select the metrics used to manage a business's operations. Additionally, KPIs can be predictive indicators – known as key predictive indicators – which can help identify when a firm is likely to miss their strategic goals before they happen (Key Predictive Indicators, 2010).

As executives and supply chain professionals are racing to adopt these solutions, it is important to note that the strategy for using and implementing these practices remains just as important as the technologies themselves. David Kiron and Michael Schrage (2019) in "Strategy for and with AI" declare that investing in artificial intelligence and machine learning is not enough. Rather, firms need to focus on leveraging these technologies to help build the strategy and implementation of these new initiatives instead of choosing technology and hoping that the competitive advantage will follow. Further, the authors point to KPIs as a specific measurement that firms need to align with their company objectives when developing these tools (Kiron & Schrage, 2019).

For example, if a large global firm were to invest in an analytical digital dashboard, their probable expectation would be that the CEO leverage this technology to create shareholder value. However, if the CEO wanted to focus on supply chain performance, he or she would likely find many different metrics used within the supply chain department. Therefore, to efficiently manage and gauge the health of the

supply chain, it would be crucial for the CEO to determine which metrics are driving shareholder value and which are not. To identify the validity of these metrics, these business leaders need a strategy to know which KPIs to focus on. For this reason, to the point of Kiron and Schrage (2019), firms need to comprehend not only how to implement a technology like a digital dashboard, but also need to understand how they can leverage these technological capabilities to select the best operational KPIs used to manage their supply chain.

Today, it is common for firms to select performance metrics for several reasons. One could be the assumptions that individuals within the industry may have about a particular business or function. For example, a firm may believe that a performance metric such as fill rate is an acceptable measure of the company's ability to deliver products to their clients based on their assumptions of customer expectation. However, the way in which companies calculate their fill rate metric can vary greatly, and the firm could be performing at a high fill rate because they are delivering all the items correctly when their customers are more interested in the speed of the delivery (Barnhart, 2012). Another reason could be that a particular metric is traditionally used within a given industry. An example of this would be lead time, which is one of the most relevant KPIs within manufacturing (Graham et al., 2015). This KPI may be commonly used amongst manufacturers, however, managing to this metric may not be the best solution for all firms within this industry. As technologies continue to evolve and data becomes more readily available for management teams, the use of KPIs becomes even more important. For these reasons, it is evident that firms should seek data-driven solutions to find the best metrics to indicate the performance of their business.

1.1. Problem Statement

This research is a student-proposed capstone that will review a data set supplied by a firm, identified as *Company XYZ.* Though *Company XYZ* tracks and measures several different metrics related to their manufacturing processes and inventory management, the question remains which of these metrics the firm should use to manage their operations and gauge the health of their supply chain. With access to thousands of data points that could provide insight, the firm is now turning its attention to finding the KPIs that are driving performance and eliminating those that are not. By down selecting the most critical performance metrics, *Company XYZ* can simplify their operations management, while reducing the confusion that can be created by insignificant measures.

Though many methods and frameworks exist to assess and rank KPIs, this research will focus on leveraging statistical modeling and machine learning practices to find which of the given metrics are the

most statistically significant considering the firm's overall objective. This approach will be consistent with standard predictive modeling techniques such as random forest and LASSO regression: selecting the outcome variable the firm aims to improve and then conducting analysis to discover which independent variables have the biggest impact on this dependent feature. This research will identify which of the manufacturing and inventory metrics provided by *Company XYZ* has the largest impact on the production efficiency KPI the firm has chosen to optimize. For confidentiality, the name of the company and description of the production efficiency metric have been concealed.

Our hypothesis is that statistical modeling will help the firm distinguish which metrics are most relevant in achieving a numerical objective and can help provide insight into the company's future performance. This analysis will identify various statistical models, cover their basic underlying methodology, and demonstrate how to use them on an example data set using the R programming language. The descriptions and examples provided can be used as a reference for professionals and academics throughout the field of supply chain management. The success of this analysis will be measured by the ability to identify the KPIs proven to have the most significant impact on the *Company XYZ's* production efficiency metric.

2. LITERATURE REVIEW

This project will determine how *Company XYZ* can select the most impactful KPIs out of a list of already established metrics to monitor its performance. This literature review will begin by identifying and describing what a KPI is and its relation to supply chain management and inventory, followed by a brief explanation of current methods used for selecting and ranking KPIs. This chapter will then cover alternative approaches to the existing ranking and selection processes using machine learning. Ultimately, this review will conclude with a clear path forward regarding the methods that can be utilized to solve the given problem statement.

2.1. Key Performance Indicators in Supply Chain Management

A key performance indicator (KPI) is an objective measurement that can compare different companies' or employees' performance under the same baseline. KPIs should link to profitability or a strategic goal such as operational efficiency to demonstrate that achieving a particular number is meaningful, as well as capable of illustrating a clear target for the firm to work toward. The key points of the KPIs are neither having too many of them nor having over-complicated measures that can mislead

those who are using them (Weinstein et al., 2019). One commonly known KPI is the net profit margin, which converts net profit to a revenue percentage (Twin, 2020). Supply chain management is one area that relies heavily on KPIs. Namely, performance measurement in inventory is an essential component in manufacturing, whereas inventory KPIs can give a hint to the health condition of the business (inventory turn, service level, item fill rate, inventory cost, etc.).

KPIs are commonly used today regardless of how new technologies have evolved. They provide a tremendously helpful guide to the areas that require attention and can detect potential underlying problems when they are selected and designed properly (Kuhfahl et al., 2018). However, as new technologies and more data become available, KPIs can be altered to better fit business needs. New ways to collect data allows additional information to be captured and thus able to develop new KPIs for measuring performance. For instance, using video sensor can collect additional data for analysis and develop additional KPIs such as more specific root causes of damage shipments.

Akyuz and Erkan (2010) compiled a literature review article on supply chain performance measurement. The article included a section that showcase different ways on modeling, prioritization, and dependence modeling of KPIs, although there are still challenges to resolve yet multiple approaches have done including analytical hierarchy process (AHP). Nevertheless, there is a limitation to solely using KPIs to track, monitor, and facilitate decisions to achieve a target. Completely aiming to achieve excellent KPI results may not drive towards the right goal.

In addition, KPIs may not always be coherent with each other; sometimes, departments have conflicting interests in achieving outstanding KPI results (Shintaro Urabe et al., 2016). For example, when launching new products, a firm's marketing team would like to offer every possible feature that customers may be interested in to fully capture the market share, whereas the firm's finance and supply chain teams focused on inventory management may reject the approach of carrying all potential product offerings. In this case, maximizing market share represented by the Marketing KPI may conflict with inventory turns, a measurement of how many times the components are used in a certain period. A lower inventory turn would result in a higher risk of excess and obsolete inventory, implying the firm is not using their capital investments effectively.

As numerous KPIs are available for organizations to choose from, the question is: what is the best way to leverage analytic techniques to select the most appropriate KPI that has the greatest impact on the underlying management focus? In the following section, we will review some of the commonly used approaches to select and predict performance metrics. By understanding these methods, we will be able

to align our predictive modeling approach for *Company XYZ* with other methodologies that have been used in prior research.

2.2. Current Methods for KPI Review and Selection

This section reviews methods for finding the performance metrics that provide the most value to the firm. Two different methods that we found throughout our review of literature are linear programming and multi-criteria decision making (MCDM). Both are used to either select or rank KPIs based on their importance to a firm's overall objectives. Given that the purpose of this research is to find how *Company XYZ* can leverage predictive analytics to assess operations metrics, these methods do not directly apply to our given problem statement. However, they provide valuable insight into how other professionals in the field have approached KPI selection, and can help us form a methodology for leveraging predictive analytics. The following sections will elaborate on MCDM – specifically data envelopment analysis and analytical hierarchy process – and linear programming.

2.2.1. Multi-Criteria Decision Making

Multi-criteria decision making (MCDM), also referred to as multi-criteria decision analysis (MCDA), is a method commonly used within operations research to help end users make an optimal decision by assessing multiple criteria (Triantaphyllou et al., 1998). MCDM can be further broken into two categories; multi-objective decision making (MODM) which is used when multiple objective functions are present, and multi-attribute decision making (MADM) which focuses on problems where the set of decision alternatives are predetermined. Through our review of literature, we found examples of two different forms of MADM being used to assess operations metrics – data envelopment analysis and analytical hierarchy process.

Data envelopment analysis (DEA) uses a linear-programming-based model to measure the input-output efficiency relationship on whether a measurable object is at best practice or not. DEA is widely used in manufacturing settings to essentially develop a best trajectory output line called a frontier. If the output performance is not on the frontier line, the output is not performing at the optimal level. This can provide management guidance on which measurable objects need improvement and in which direction. Liu and Liu (2008) conducted a case study with a manufacturing company that has multiple input-output indices on its nine production lines. They successfully used DEA to evaluate which production line is not at an optimal level and how it can be improved by scaling up or down on the production activities. This technique measures and evaluates the performance of specific outputs based on various input variables

and focuses on how much the output could be improved rather than determine the relative importance of specific input variables.

The DEA method works well to evaluate efficiency performance and recognize areas for improvement. However, it is not suitable for identifying the importance of individual variables or selecting and ranking the input measurement that would have the greatest impact on the output KPI as the method blends all variables together to create a formulation. There are also hybrid variations of DEA combined with different theories. However, those theories are still based on DEA, which is not relevant to ranking and selecting the key performance metrics that have the greatest impact on the objective function.

Analytical hierarchy process (AHP) is the most used MCDM method, with thousands of papers published on different applications for this method (Munier & Hontoria, 2021). AHP is a decision-making framework, which allows users to decompose problems into hierarchal levels (Schmidt et al. 2015). Further, weights for each criterion and its alternative are assigned based on pairwise comparisons, and the solution is then found utilizing eigenvector calculations. Through our review of literature, we found an example of AHP being used to select KPIs for occupational safety and health management systems (OSH MS). Researcher Daniel Podgorski (2013) employed this method to create a pairwise comparison of each of the various metrics and KPI categories his team identified to create a vector with each entity having an assigned weight based on a certain set of criteria they labeled as SMART (specific, measurable, achievable, relevant, time-bound). The result of Podgorski's research was a ranked list of performance metrics for each of the 20 categories (Podgorski, 2013). This research showed the validity of a rank-based approach to selecting KPIs; however, given the limiting constraints of the provided dataset, methods outside of AHP need to be used to properly rank the importance of the various manufacturing and inventory management metrics in our research.

2.2.2. Linear Programming

Linear programming is a commonly used technique to find the optimal solution to an objective function, given various linear constraints (Hayes, Pakornrat, & Khim, 2020). Given that *Company XYZ* has selected a production efficiency metric that they aim to maximize, this method was identified as a potential solution to the given problem. Different forms of linear programming include integer linear programming (ILP) where only integers can be considered for the function's decision variables, and mixed-integer linear programming (MILP) where only some decision variables are integers. Stricker, Minguillon, and Lanza (2017) used an ILP method to select KPI metrics within a production system of regulation valves. Their approach focused on minimizing the number of metrics needed to meet the information

requirements they had set forth. Their team identified that KPI selection is often subjective in nature and chose this process because it relies on a mathematical formula to select the amount of KPIs chosen, whereas other approaches such as analytical hierarchy process (AHP) rely more heavily on subjective measures (Stricker, Minguillon, & Lanza, 2017). KPIs were then measured and selected based on three characteristics: sensitivity, explanatory power, and interval divergence. Though these findings were mathematically obtained, the necessity of a rigorously constructed dataset and specialized software makes this method unsuitable for the project, given the problem and the dataset. If *Company XYZ* were to select this approach for KPI selection, there would need to be a sizable investment in time and assets.

2.3. Machine Learning Models for Feature Selection and Ranking

Several methods to select and rank KPIs have been reviewed such as MCDM and linear programming. Though these methods have been proven within the field of operations research, our analysis aims to find a new alternative in utilizing statistical modeling to rank and select performance metrics without the rigor of the traditional approaches. The dataset provided by *Company XYZ* is structured like a statistical modeling problem, with a dependent variable (the firm's production efficiency metric) and several independent variables (manufacturing and inventory KPIs). Given the structure of the problem, our research will focus on applying statistical modeling techniques rather than the previously stated methods.

Though using machine learning and predictive modeling is not yet commonly used for KPI selection, Nenad Stefanovic (2014) published a journal that outlines the potential of using these methods for predicting future supply chain performance using a firm's KPIs. Stefanovic outlines decision tree and regression algorithms as ways to down select variables and predict performance, and his research is concluded with a dashboard application which identifies key metrics and predicts future performance (Stefanovic, 2014). Our research will apply a similar methodology to provide the *Company XYZ* with a fast and cost-effective method to analyze their already established performance metrics.

The models described above tackled the problem of selecting KPIs in two different ways: ranking variables based on their importance to an objective measure, and selecting variables based on their sensitivity and explanatory power to an objective measure. Research shows that these two methods (ranking and selecting variables) can be executed utilizing multiple linear regression, random forest, backward elimination, and LASSO regression.

2.3.1. Multiple Linear Regression

Multiple linear regression (MLR) is a commonly used statistical modeling tool. As defined by Will Kenton, MLR "is a statistical technique that uses several explanatory variables to predict the outcome of a response variable" (Kenton, 2020, Multiple Linear Regression). From this definition, the explanatory variables would be the independent variables, and the response variable would be the dependent variable. Commonly used when the number of observations is larger than the number of variables (Grömping, 2009), MLR creates a prediction for the dependent variable based on the values of the independent variables. MLR provides a coefficient and p-value for each independent variable, allowing users to easily identify the significance of a given value in predicting the outcome of the response variable (p-value), and the impact that the explanatory variable has on the response variable as it increases or decreases in measure (coefficient). With its easy interpretability, MLR is commonly used by professionals across industries. However, there are limitations to this model. For example, MLR assumes that the independent variables are not too highly correlated, and that there is a linear relationship between dependent and independent variables (Kenton, 2020). Therefore, proper modeling techniques such as checking for linearity and multicollinearity must be taken when using this approach.

As stated by Grömping (2009), there are several ways to measure the importance of an independent variable within a MLR model. However, software libraries such as the caret package in R provide a simplistic way to find this measure. Using the method available in caret, each parameter is ranked based on the absolute value of the t-statistic (R Package Documentation, 2020). This method will provide a ranking of each independent variable, which can be contrasted with the results of the variable rankings provided by other statistical models.

2.3.2. Random Forest

Random Forest is another popular machine learning model, which can evaluate datasets containing a wide range of observations and variables while providing an assessment of variable importance (Grömping, 2009). Unlike MLR, Random Forest models allow for nonlinearities to be assessed without the need to explicitly model them. This model is tree-based, and therefore reliant on regression trees that recursively partition data, ultimately creating a regression function built on a multidimensional step function (Grömping, 2009).

For ranking Random Forest variables, Grömping (2009) discusses two different approaches. The first variable importance measure reviewed was the Gini importance approach, which ranks features according to their average impurity reduction when variables are split. Though this is a commonly used form of ranking variables, the variable importance metrics can become biased due to some split variables (Grömping, 2009). The second method is MSE reduction, which ranks variables according to their predictive accuracy. Both measures can be found utilizing the Random Forest package available in R (R Package Documentation, 2020).

2.3.3. Backward Elimination

With two separate approaches identified to rank KPIs, our research will now turn to methods for selecting these metrics. Through the review of variable selection methods conducted by Heinze, Wallisch, and Dunkler (2018), backward elimination was identified as one of the most common feature selection algorithms used today. Backward elimination is similar to another popular step function algorithm – forward selection. Both methods recursively regress a set of independent variables against the dependent variable, until only independent variables found to be statistically significant remain. Backward elimination begins by taking all the independent variables given in a model, removes the least significant in predicting the dependent variable, refits the model with the remaining independent variables, and repeats this process until no insignificant independent variables remain (Heinze, Wallisch, & Dunkler, 2018). Forward selection begins by building a univariable model for each independent variable, selecting the univariable model with the highest level of significance, regressing the remaining independent variables and selecting the most significant independent variables to add to the model until there are no more significant independent variables to include. Though both these approaches can select significant variables, many statisticians prefer the backward elimination model when collinearity is present between the independent variables (Mantel, 1970).

Backward elimination is also an enticing feature selection algorithm because it works as an extension for multiple linear regression models. As stated previously, MLR is one of the most common statistical modeling techniques used today. With its individually assigned p-values and coefficients, these models provide analysts and business professionals with transparency into how each independent variable impacts the prediction of a dependent variable. With backward elimination algorithms available in numerous open-source R packages (R Package Documentation, 2020), this feature selection technique will be simple, cost-effective, and interpretable for *Company XYZ*.

2.3.4. Least Angle Selection and Shrinkage Operator (LASSO)

Another feature selection algorithm reviewed by Heinze, Wallisch, and Dunkler (2018) was least angle selection and shrinkage operator, often referred to as LASSO. LASSO is most commonly used with high-dimensional datasets where the number of independent variables exceeds the number of individual records, however this technique could be utilized on any dataset of at least ten independent variables (Heinze, Wallisch, & Dunkler, 2018).

The LASSO method selects features for a model through regularization – a shrinkage process that penalizes independent variables' regression coefficients down to zero (Fonti, 2017). The variables at the end of the shrinkage process that have a non-zero coefficient are those which are selected for the model. Regularization is a powerful concept that minimizes prediction error and reduces the likelihood of an overfit model through the reduction of collinearity between selected independent variables (Nathan, 2019). Another common regularization method not covered in detail by Heinze, Wallisch, and Dunkler (2018) is ridge regression. Like LASSO, ridge regression utilizes a shrinkage method, however it only reduces regression coefficients and does not eliminate them (Fonti, 2017). For this reason, our analysis for *Company XYZ* will use LASSO as a preferred regularization choice. Like the other statistical modeling techniques reviewed in this analysis, LASSO is available within R using the glmnet library (R Package Documentation, 2020).

2.4. Literature Review Summary

Company XYZ has provided a dataset of 10 independent KPIs with one dependent KPI, which they are trying to maximize. Our review has covered what KPIs are and their importance in business and supply chain management, as well as commonly used methods to rank and select KPIs. Our analysis has identified multiple linear regression and random forest as methods to rank these KPIs, and backward elimination and LASSO as methods to select and eliminate KPIs within the given dataset. This research will now turn to applying these machine learning methods utilizing the R programming language to solve the given problem statement.

3. DATA AND METHODOLOGY

This research begins by reviewing the dataset provided by *Company XYZ* and discusses the processes to clean and validate the data. We then review in detail the methodology that will be used for the four different modeling approaches, and finally discuss how we will compare our findings for each

model. This section will provide a clear path forward to solve the problem statement of identifying the best method to select KPIs utilizing machine learning and statistical modeling.

The dataset provided by *Company XYZ* describes the production of a product over an undisclosed period. It consists of eighty-eight records, each representing an individual line item that was manufactured in sequential order. To further explain, imagine a hypothetical manufacturing plant that creates make-to-order products. This dataset would represent a manufacturing line that creates a single type of product, with the first record representing the first product created in the time-period and the last record representing the last product created in the time-period.

The dataset also has twelve columns; one column identifies the line item (e.g., an individual product), one column represents the dependent variable that the firm aims to maximize (e.g., a production efficiency metric), and the other ten columns represent the independent variables which are various production KPIs that were recorded for each record. The names of the KPIs are not disclosed in order to keep *Company XYZ* and its manufacturing processes anonymous. However, the KPIs associated with manufacturing have been labeled with an 'X' ($X_1, X_2, \dots, X_7$), and the KPIs associated with inventory management have been identified with an 'I' (I_1, I_2, I_3). In total, the dataset provides ten independent variables – seven manufacturing KPIs and three inventory management KPIs.

An important caveat to recognize with this dataset is that it was cleaned and scaled by *Company XYZ* before our research team received it. This means that outlier records were removed from the dataset, and all variables were normalized on a scale of zero to one. Even though this allows us to omit these steps before analyzing the data, our analysis still briefly covers the processes of removing outliers and scaling data as it is an integral part of any statistical modeling application.

3.1. Exploratory Data Analysis (EDA)

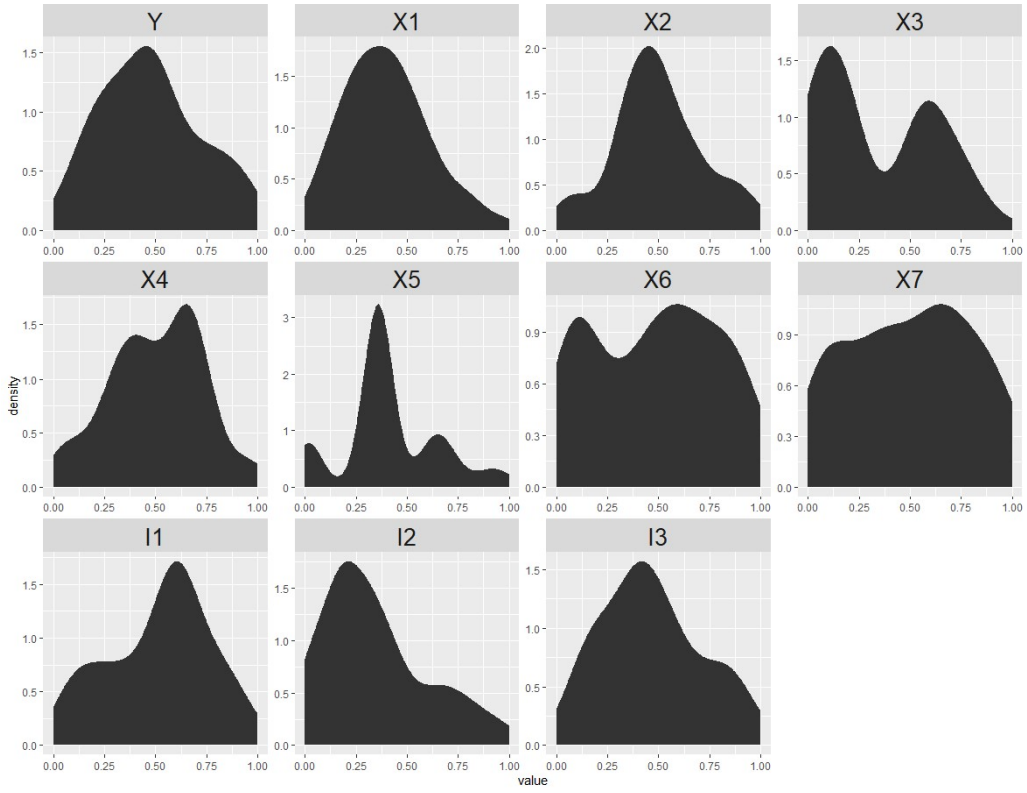
Before conducting statistical modeling on a dataset, it is important to thoroughly analyze the data points to find anomalies and patterns that could affect the model's output (Patil, 2018). This pre-modeling process is commonly referred to as exploratory data analysis (EDA). In this section, our analysis will review the steps needed to perform EDA on our dataset to identify the distribution and correlation of the variables, as well as discuss the process to remove outliers and normalize the data points.

3.1.1. Distribution of Variables

This analysis explores the distribution of dependent and independent variables in the dataset. *Figure 1* shows the results of a histogram which plots the distribution of the 11 variables within our dataset. The histograms in this display provide a visual representation of each variable's distribution, allowing our research team to determine whether these variables are normally distributed. Though normality in all variables is not a requirement for models such as multiple linear regression (Kim, 2015), it is still necessary that residuals are normally distributed for all predictive models. This analysis can also provide valuable insights into the dataset, such as relationships between variables. *Figure 1* demonstrates that most of the variables are normally distributed, except for variables X_3 and I_2, which are skewed slightly right. Knowing that the data was normalized and outliers were removed beforehand, these findings validated that there is no need for further manipulation before modeling due to variable distribution.

Figure 1

Variable Histograms



3.1.2. Correlation of Variables

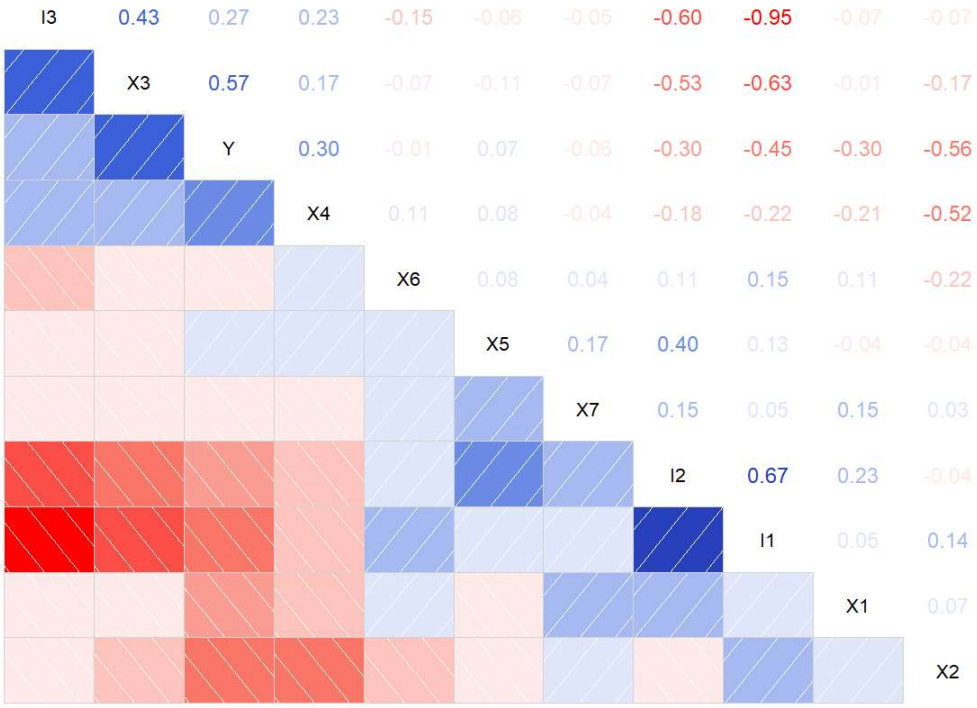
Another technique that can help provide insights into the variables for a dataset is checking the correlation between these variables. Once two or more independent variables are highly correlated, they can lead to a statistical occurrence known as multicollinearity – which is the presence of highly correlated and redundant predictors (Hayes, 2020). Multicollinearity leads to various issues in a statistical model – including overfitting, unstable coefficient estimates, and difficulty identifying significant variables (Wu, 2020).

Understanding that our dataset consists of manufacturing KPIs, we anticipate that there will be some correlation between the independent variables. For example, if a line item had a strong performance in one inventory KPI, one could suspect that the line item would also perform strongly in one of the other two inventory KPIs. Due to the limitations of the size of the dataset, our analysis will not exclude any independent variables from our dataset at this time. Instead, we will derive insights from this process and check the variance inflation factor (VIF) of the independent variables selected for our final MLR models.

VIF is the reciprocal of variable tolerance, which is defined as the proportion of variance an independent variable shares another independent variable in the analysis (O'Brien, 2007). Though there are no set standards for accepting or rejecting variables with a particular VIF, there are common rules of thumb for this metric such that a VIF score higher than ten often indicates excessive levels of multicollinearity (O'Brien, 2007). If highly correlated variables are identified using VIF, our team will have the option to eliminate variables from the model or consolidate the two measures. *Figure 2* shows a correlation matrix for all the variables within our dataset. The color coding represents how correlated two variables are, with a dark blue square representing a strong positive correlation, and a dark red square representing a strong negative correlation.

Figure 2

Correlation Matrix



Focusing on *Figure 2*, we can draw a few insights into our variables. As anticipated, we observe a correlation between the three different inventory KPIs. Variables I_1 and I_2 have a strong positive correlation, while I_3 has a strong negative correlation with these two variables. It is also evident that there is a correlation between our dependent variable Y and the independent variables I_1, X_2, and X_3. As stated previously, we will not remove or alter any variables from our dataset at this step but will continue to monitor potential multicollinearity within the models using VIF and consolidate and manipulate independent variables as needed.

3.1.3. Removing Outliers

Though outliers were already removed from the given dataset, we will briefly cover the importance of this process. Outliers in a dataset could arise if an anomaly occurred for a particular record or a mistake was made in the collection of the data. With little insight into what the independent variables of the dataset are or the data collection process, our team would not be able to distinguish the difference between a potential anomaly in performance or an error in data collection. For example, extremely poor performance for one particular line-item would potentially be identified as an outlier, whereas a ranking

above a one for a metric that ranges between zero and one would likely be a mistake in the data collection. Researchers may want to keep anomalies within a dataset to represent variability, while faulty data should always be removed. However, understanding that outliers were removed from this dataset, this identification process is not necessary to conduct further analysis.

A common practice when removing outliers from a dataset is to rely on mathematical functions such as z-scores or interquartile range (IQR) to identify and remove these records. These mathematical functions identify the distribution of the data and then eliminate records that are found outside of a threshold given by the analyst (Sharma, 2018). Removing these outliers will decrease the variability that the statistical models can identify between variables. However, it also creates a dataset less biased from anomalies and potential data collection mishaps – ultimately leading to a stronger predictive model. Our analysis confirmed that outliers were not present in the dataset when the distribution of all variables was plotted (refer to *3.1.1. Distribution of Variables*).

3.1.4. Scaling Variables

The dataset used for this analysis was scaled and normalized by *Company XYZ*. Though this was stated in the disclaimer given within the data file, it was also evident when our analysis concluded that each variable had a range between zero and one (refer to Figure 1). This process is important when conducting statistical analysis when numerical columns have differing ranges of values (Lakshmanan, 2019).

For the example of manufacturing and inventory KPIs, it is logical to conclude that these metrics could have vastly different scales. For example, one metric may be a ratio that results in a value between zero and one, while another metric could be a count of instances that the company chooses to measure. In this scenario, if the ratio variable had values between zero and one and the count variable ranged between one hundred and one thousand, the variable with the larger values would intrinsically influence the results of the data model. By scaling all variables to have values within the same range, the potential for bias between variables in the model is removed without distorting the variance between each of the variables (Lakshmanan, 2019).

3.2. Modeling

We will now turn our attention to applying the four machine learning techniques identified in the literature review. These modeling techniques are broken into two categories: variable selection and variable ranking. For variable selection, our analysis covers backward elimination and LASSO. These

methods will be conducted on our dataset and will result in a subset of independent variables that the model identifies as best fit to predict the outcome of our dependent variable.

Following this, linear regression and random forest methods will be used to rank the importance of each independent variable in our dataset. Like the AHP method discussed in the literature review, we will discern each of the variables into two categories – manufacturing and inventory metrics. Using the variable rankings obtained by each approach, our analysis will select the one variable found to be most important in the inventory metric category and three variables found to be most important in the manufacturing metric category. Ideally, our team would break seven manufacturing metrics into three different categories and select the most important KPI for each. However, due to the lack of information for the names and descriptions of these KPIs, we will rely on the assumption that at least three of these manufacturing KPIs would be needed to provide sufficient insight into the operations of the manufacturing plant.

3.2.1. Backward Elimination

As described in the literature review, backward elimination works by assessing all the independent variables given in a model and recursively removes the least significant until no insignificant independent variables remain (Heinze, Wallisch, & Dunkler, 2018). This method was chosen for its interpretability and its ability to outperform forward selection when collinearity is present between the independent variables (Mantel, 1970).

This technique will be carried out by first running a linear regression model, and then backfitting the linear regression model. Understanding the potential for multicollinearity between the inventory variables, we will then check the VIF of the selected independent variables to assure they are not highly correlated. *Table 1* displays the results of the backward elimination model.

Table 1

Backward Elimination (Preliminary Results)

Variable	Coefficient	t-value	p-value	VIF
Intercept	2.42719	13.874	<2e^-16	
X1	-0.31593	-4.858	5.6e^-6	1.014325
X2	-0.42017	-6.576	4.2e^-09	1.082538
X5	0.19791	3.287	0.00149	1.069239
I1	-1.85849	-10.111	4.51e^-16	11.417948
I3	-1.56328	-8.479	7.82e^-13	11.109632
Median Residual		0.005		
R2		0.7469		
Adjusted R2		0.7314		

Based on these preliminary results, it is evident that the backward elimination method has selected a group of independent variables that have a strong ability to predict the outcome of our dependent variable. All independent variables selected are found to be statistically significant (p-value < $\boldsymbol{\alpha}$ – which our team has set at .05), the residuals are unbiased (median residual 0.005), and an adjusted R-squared of 0.7314. However, it is important to note that the VIF score of the inventory variables (I_1 & I_2) is above 10, meaning that they are highly correlated and likely affecting our regression coefficients. In chapter *4.2. Addressing Multicollinearity,* we discuss how we can remove these highly correlated variables to create a more robust final model.

3.2.2. LASSO

Another feature selection method selected for this analysis is LASSO. This method was selected for its use of regularization, which is a process that penalizes and shrinks the value of an independent variable's coefficient down to zero (Fonti, 2017). This method eliminates variables it finds to be insignificant in predicting the outcome of the dependent feature, or in this case, the production efficiency metric. Though similar to other methods identified in the literature review such as ridge regression, LASSO was ultimately chosen for utilizing both regularization and feature elimination.

Once these data structures for this model are created, our team can then proceed with the analysis. For this model, we will use the lambda parameter of 'lambda.1se'. Lambda is the regularization parameter within a LASSO model, and 'lambda.1se' will create the most regularized model where the error is within one standard error of the minimum (Hastie & Qian, 2014). Table 2 shows the preliminary results for the LASSO model.

Table 2

LASSO (Preliminary Results)

Intercept	1.1254
X1	-0.2732
X2	-0.4916
X3	0.2823
X4	-0.0533
X5	0.0967
X6	-0.0099
X7	.
I1	-0.4706
I2	.
I3	-0.3328

These preliminary results show that the LASSO method has selected eight of the ten potential variables while eliminating X_7 and I_2 from the variable set. Although the coefficients differ from those given in the backward elimination model, it is encouraging to observe that the variables both models have in common affect the dependent variable in the same way, meaning that both models recognize the coefficients to be positive or negative. Another similarity to the backward elimination model is that both the I_1 and I_3 variables were selected. Therefore, as a method for consolidating or eliminating these correlated inventory metrics is identified, the results of this model are likely to change.

3.2.3. Multiple Linear Regression

When reviewing methods to rank variables and their importance in predicting the outcome of a dependent variable, our analysis found MLR as an ideal modeling solution for our problem statement. MLR is widely used by professionals and creates an easily interpretable model output which assigns p-values and coefficients to each independent variable (Grömping, 2009).

For this modeling exercise, we will use an MLR model to regress all ten independent variables against the dependent variable. Once this model is performed, we identify the ranking of variable importance based on the absolute value of the t-statistic. As explained prior, our team will keep the three highest-rated manufacturing KPIs and one of the inventory KPIs. *Table 3* lists the results of the MLR model, and *Table 4* lists the variables by their predictive importance in descending order.

Table 3

Multiple Linear Regression (Preliminary Results)

Variable	Coefficient	t-value	p-value	VIF
Intercept	2.41496	8.241	3.49e^-12	
X1	-0.30717	-4.233	6.31e^-5	1.208852
X2	-0.47755	-5.774	1.55e^-7	1.736458
X3	0.01736	0.189	0.85051	3.421554
X4	-0.07186	-0.929	0.35567	1.619461
X5	0.21959	3.077	0.00289	1.437369
X6	-0.03007	-0.624	0.53446	1.124575
X7	-0.01474	-0.306	0.7601	1.068982
I1	-1.74582	-5.676	2.33e^-7	30.60586
I2	-0.04421	-0.486	0.62827	2.81682
I3	-1.48416	-5.557	3.78e^-7	22.31044
Median Residual		0.00359		
R2		0.7516		
Adjusted R2		0.7194		

Table 4

MLR Variable Importance

Rank	Variable	Overall
1	X2	5.7737847
2	I1	5.6755978
3	I3	5.5574559
4	X1	4.2330174
5	X5	3.0774629
6	X4	0.9292372
7	X6	0.624019
8	I2	0.4861064
9	X7	0.3064355
10	X3	0.1891091

Based on the roles of categories mentioned previously, the preliminary results shown in *Table 3* and *Table 4* of this model would select variables X_2, I_1, X_1, and X_5 as the most important variables. We are not considering variable I_3 as we are only looking to select one variable in the inventory metric category. Like the previously reviewed models, the multicollinearity between I_1 and I_3 in this model is evident through the VIF metric, and when these variables are consolidated or deleted, we are likely to see slightly different results.

3.2.4. Random Forest

With the ability to evaluate datasets of various sizes and model nonlinearities, Random Forest is a popular machine learning algorithm capable of ranking variables by their importance (Grömping, 2009). Utilizing a tree-based model that creates a regression function on a multidimensional step function, Random Forest has two commonly used ways to rank variables – Gini importance and MSE reduction. As explained in the literature review, we will focus on the variable importance metrics produced by the MSE reduction approach, as there are known biases that can be created using the Gini importance method (Grömping, 2009).

Unlike MLR, Random Forest has parameters that can be set to best fit the model. Utilizing the tuneRF function in the RandomForest package in R (R Package Documentation, 2020), our team was able to identify the optimal mtry parameter for our dataset to be six. For further explanation, mtry is the

number of variables available for splitting at each tree node (Brownlee, 2016). By changing the mtry from the default setting of the square root of the number of columns, our model is able to minimize the OOB error (refer to *Figure 3*). Without a readily available package to optimize the ntree parameter, which is the number of trees for the model to consider, our team attempted to optimize the results through trial and error. After numerous attempts, we found little impact in the variable importance the model was predicting, and therefore decided to set the ntree parameter to the default of 500.

Figure 3

mtry Optimization

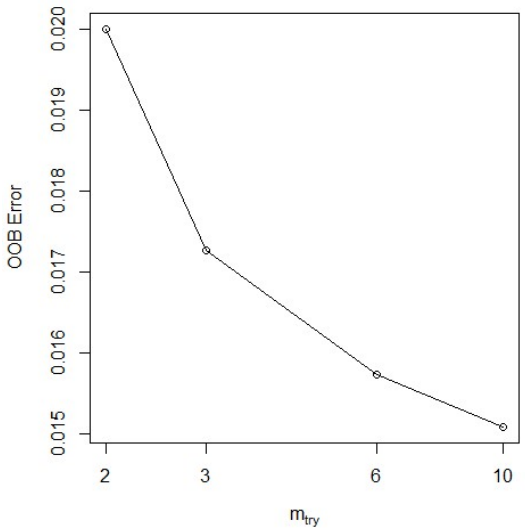
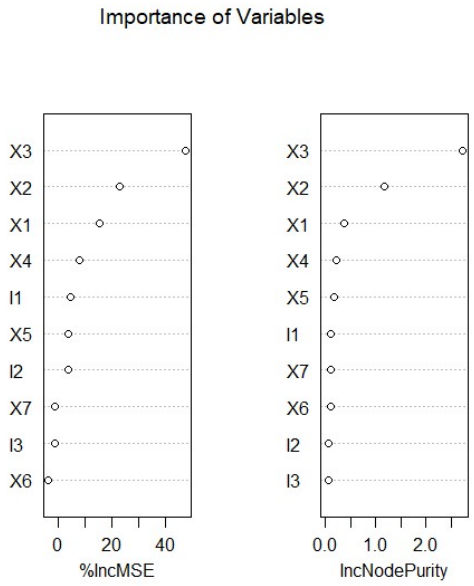


Figure 4

Random Forest (Preliminary Results)



Based on roles of inventory and manufacturing categories, the preliminary results of this model would select variables X_3, X_2, X_1, and I_1 as the most important variables (refer to *Figure 4*). If our team were to use the MSE reduction method (%incMSE) or the Gini importance method (IncNodePurity) to rank variables, the results of the selected variables for this exercise would not change. As previously stated, the importance of the inventory metrics is subject to change after a decision is made to consolidate or remove them.

3.3. Sensitivity Analysis

Sensitivity analysis is a method to measure the uncertainty of the output variable by changing the input variables (*Sensitivity Analysis in Linear Regression*, 1988). This analysis provides an opportunity to evaluate and recognize how these variables interact with each other as a linear regression model.

Sensitivity analysis can also lead to the development of different scenarios analyzing the underlying situation when input coefficient changes. Sensitivity analysis is widely used across multiple industries and disciplines. A good modeling cannot go far without sensitivity analysis to ensure the model effectively captured all the significant variables.

One application of sensitivity analysis is to observe how the predicting variable varies along with the changes in independent variables. Abdallah et al. (2012) developed a closed-loop location inventory model following sensitivity analysis where the analysis showed how individual variables impact the decision of number of facilities open. Abdallah et al. (2012) found that over a certain value of the independent variable, the model suggested the output become economically infeasible. This provides great insight to management teams when making decisions. Similar approaches can be applied to our final model.

We will first evaluate how the selected KPIs impact our predicting output variable. This can be done by interpreting the linear regression coefficients. Second, with an optimization and simulation model, *Company XYZ* can simulate different scenarios based on potential events and can see how these events interfere with the independent variables, and thus how the predicting variables interact along with the efficiency metric. With these insights, *Company XYZ* can better plan and react to the potential events by developing a mitigation plan. Furthermore, *Company XYZ* can simulate with different variable inputs to understand whether certain combinations of the variable coefficients are feasible.

Overall, the sensitivity analysis will show how the output variable fluctuates when one or many of the input variables change. From here, companies can draw scenarios and simulations to better plan for the future, maintaining a desired dependent output KPI.

3.4. Methodology Summary

In this chapter we reviewed *Company XYZ's* dataset, how we manipulated the data, and how we applied the four different modeling techniques highlighted in the literature review. By applying this methodology, we were able to observe which variables and features were best at predicting the firm's desired outcome and addressed how these models can be used to conduct sensitivity analysis relating to the metrics within the model. In the next chapter, we will identify the best subset of KPIs for predicting the dependent variable using multiple linear regression and showcase an example of how the firm could leverage the model's sensitivity analysis.

4. RESULTS

This chapter begins by reviewing the results of the four different models covered in the methodology. Our objective was to choose independent variables for a final predictive model that are accurate, concise, and easily interpretable. These characteristics were important to the model for numerous reasons. First, the model must be capable of predicting the firm's efficiency metric. With an accurate linear regression model, our team can utilize the model's coefficients to build a strong sensitivity analysis that will help the firm plan for different scenarios moving forward. With a concise subset of metrics, we can eliminate KPIs that do not impact the dependent variable. Finally, by having an interpretable model the firm will have clarity into which metrics are have the largest impact on their production goals, and the targets that must be set for each metric to attain these goals.

Once the selected KPIs and final model are identified, we will review our sensitivity analysis. Using the multiple linear regression coefficients as the building-blocks for this analysis, we will discuss how this sensitivity analysis can be utilized by *Company XYZ* and review a scenario in which the performance represented by the inventory metric I_1 is expected to depreciate by 10%. For this scenario, we use an optimization model to show how much better the firm must perform with their manufacturing metrics to maintain or even improve their production efficiency metric.

4.1. Comparing the Models

Our analysis now turns to comparing the results of the four models. For this step, we will take the features selected by each method, and run four independent linear regression models for each of them. Our team will then assess the mean absolute error (MAE) of each model to determine which set of KPIs has the highest level of accuracy in predicting the production efficiency metric. Before selecting MAE as the forecasting metric to measure the models, our team reviewed other commonly used metrics for forecast accuracy such as Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). MAPE is the sum of the individual absolute errors divided by the prediction for each given period (Vandeput, 2019). RMSE is the squared root of the average squared error. Though both these measures would be sufficient in calculating the accuracy of our predictions, our team ultimately decided to use MAE because of its simplicity and equal treatment of all residuals (Vandeput, 2019). *Table 5* depicts the KPIs selected by each method.

Table 5

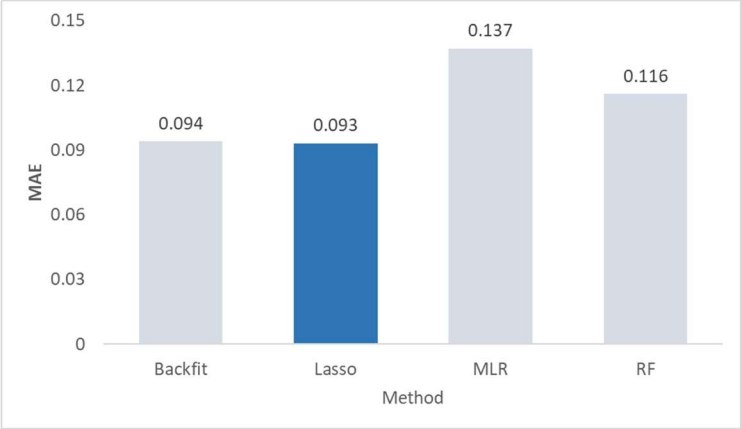
Summary of Preliminary Results

Model	KPIs Selected
Backward Elimination	X1, X2, X5, I1, I3
LASSO	X1, X2, X3, X4, X5, X6, I1, I3
MLR	X1, X2, X5, I1
Random Forest	X1, X2, X3, I1

Now that our variable sets are identified, the four different predictive models can now be created. Once each model is built, our team will then calculate the MAE of each model to find which method of selecting KPIs had the least error in predicting the production efficiency KPI. The MAE for each of the four models is listed in *Figure 5*.

Figure 5

Preliminary Results Summary



From these preliminary results, a few conclusions can be drawn. First, the selection models (backfit & LASSO) created models with ~26% lower MAE than the feature ranking methods (MLR & RF). Even though it may appear that one methodology is superior in identifying the best set of KPI metrics to predict an objective, these results are likely skewed by multicollinearity. Considering the ranking methods, our analysis has selected one inventory metric only, variable I_1, which was found to have the greatest importance in both MLR and Random Forest. For the selection methods, however, these algorithms selected the variables which resulted in the highest level of predictability. Therefore, the selection models selected the two variables (I_1 and I_3) that were identified during the EDA process as being highly correlated, as they would likely create an overfit model (Wu, 2020). Given that we tested our preliminary

models on the same dataset that was used to fit them, the likelihood of the models created for the selection methods being overfitted is high.

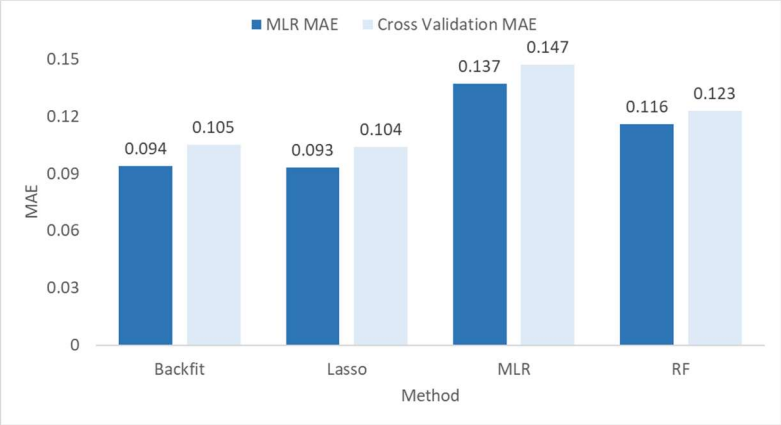
4.1.1. Cross Validation

As a method to validate and test the skill of the four linear regression models, our team conducted repeated k-fold cross-validation for each model. This method is a commonly used technique to test machine learning models. K-fold cross-validation works by splitting the given dataset into subsets (identified as k-folds), withholds one of the subsets and trains the model, and then tests the trained model on the previously withheld subset. By using the repeated k-fold cross-validation method, this same approach is applied a select number of times (identified as repeats) over randomly selected subsets, with the final model error taken as the mean of each repeat (Kassambara, 2018). This method is robust because of its ability to test the models against multiple randomly generated test datasets, which is why our research team chose it.

One important step before conducting this validation method was to select the number of k-folds and repeats needed. Through our research, we found that 5 or 10 k-folds are often selected for this analysis, as these values have been shown to yield results that do not suffer from excessively high bias or variance (James et al. 2014). Understanding that we have a relatively small number of data records, our team chose 5 k-folds as it is the smaller of the two alternatives. For repeats, we found that no more than ten repeats are often sufficient for the cross-validation process (Brownlee, 2018); therefore, we set this parameter to 10. The repeated k-fold cross-validation was applied to each of our four models and the results of the MAE for each model can be found in *Figure 6*.

Figure 6

Cross Validation MAE Comparison



From *Figure 6*, we observe small increases in MAE across all four of the models. This was anticipated, as models will typically lose some level of prediction accuracy as new data is evaluated. However, through further calculation we identified that the MAE for the backward elimination and LASSO models (which contain the two correlated variables) increased ~12% while the multiple linear regression and random forest models only realized an increase of ~6.5%. As identified in previous sections, the models containing variables I_1 and I_3 are likely overfit due to multicollinearity, and the results of our cross-validation models also suggest this theory. In conclusion, though multicollinearity has been identified as a point of concern, the evidence of our cross-validation analysis suggests that all four methods selected KPIs that can predict the outcome of the production efficiency metric.

4.2. Addressing Multicollinearity

Understanding that multicollinearity was present between the I_1 and I_3 independent variables, we reconducted analysis to see how the outputs of the models would change if one of the correlated variables were removed from the dataset. Applying the previously used methodology for all four models, the results of removing the I_1 and then I_3 variables are listed in *Table 6* and *Table 7*.

Table 6

Model Results Excluding Variable I_3

Model	KPIs Selected
Backward Elimination	X1, X2, X3, X4, X5, I1
LASSO	X1, X2, X3, I1
MLR	X1, X2, X3, I1
Random Forest	X1, X2, X3, I1

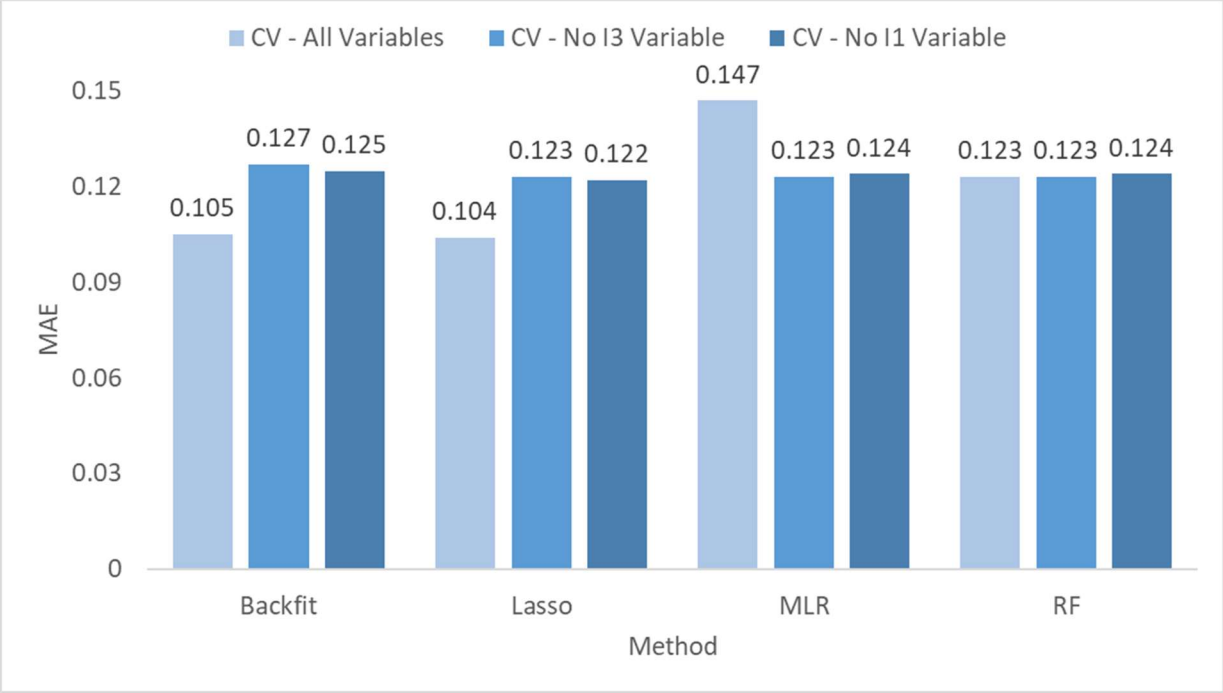
Table 7

Model Results Excluding Variable I_1

Model	KPIs Selected
Backward Elimination	X1, X2, X3, X4, X5
LASSO	X1, X2, X3
MLR	X1, X2, X3, I2
Random Forest	X1, X2, X3, I2

Figure 7

Cross Validated Results of Models When Excluding One of the Correlated Variables



The hypothesis that these models were affected by multicollinearity is confirmed in *Figure 7*. When all variables are available for the model, the selection algorithms keep the two highly correlated variables and overfits the models – resulting in a low MAE. The multicollinearity also affects the MLR model when all variables are present, by impacting the coefficients and variable importance of the different features – leading to a less accurate model. When one of the two highly correlated variables is removed, all four models recommend similar subsets of KPI metrics with similar predictive performance.

The LASSO algorithm produced the lowest cross-validated MAE when the I_1 variable was removed from the dataset, resulting in a KPI subset of X_1, X_2, and X_3. These findings suggest that our model is most accurate in predicting the outcome of the dependent variable when none of the inventory KPIs are considered. However, the KPI subsets that did include an inventory metric had a cross-validated MAE only 0.001-0.002 higher. This once again highlights the differences between the selection and ranking methods and raises the question of whether a metric should be considered if it is not found to be statistically significant in predicting a firm's desired outcome.

4.2.1. Factor Analysis

Next, we reviewed Factor Analysis (FA) as a potential method to create factors out of the 10 independent variables within our dataset to capture the variability of the different metrics while removing the potential for multicollinearity. FA models variables as linear functions of the "factors" (Penn State, Eberly College of Science). For this reason, we chose FA over the other common data reduction technique, Principal Component Analysis (PCA) – which creates new variables that are linear combinations of the observed variables, resulting in often unclear interpretation of what variables are affecting the model.

We began FA by breaking our variables into their two respective categories, which is manufacturing and inventory metrics. Then we used various statistical tests to see if the independent variables had the required characteristics for this analysis. The Kaiser-Meyer-Olkin (KMO) test is used to measure the sampling adequacy (MSA) for each of the variables and the overall model – needing to result in a MSA score greater than .5 to meet the requirements for FA (Effendi, Khairani, & Adnan, 2019). Next, the Bartlett's test of sphericity estimates whether the correlation between the variables is statistically different from zero and is required to be found significant ($p < 0.001$). The results of these two tests both passed the thresholds required to conduct FA for the manufacturing and inventory subsets. The manufacturing variables had a MSA score of 0.55 for KMO and a p-value of less than .05 for the Bartlett's test, and the inventory variables had a MSA score of 0.63 for KMO and a p-value of less than .05 for the Bartlett's test.

Once the data has been found sufficient for this analysis, we began building our factor using the factanal function in the R core package (R Package Documentation, 2020). This function performs the maximum likelihood estimation, which according to Clay Ford (2016) is often the most preferred method for FA. Looking to create only a single factor, we chose the default rotation parameter of "varimax" which keeps the rotated axis perpendicular (Ford, 2016). Figures 8 and 9 show a graphical representation of the loading pattern for the two generated factors. In FA, the independent variables used to generate a factor are each given a measurement referred to as 'loading' ranging from -1 to 1 (Ford, 2016). A loading factor close to 0 has a weak influence on the newly generated factor, while the loadings close to 1 or -1 have highest levels of influence.

Our analysis showed that the inventory factor used all three of the inventory metrics, with each of these variables having a high level of influence on the factor (refer to *Figure 8*). However, the factor created for the manufacturing metrics displays loadings for only two of the variables – X_4 and X_2 (refer to *Figure 9*). This was driven by the fact that the loadings associated with the other variables all fall

between -0.2 and 0.2, meaning that each of the other factors had little to no influence in creating the manufacturing factor.

Figure 8

Factor Loading (Inventory)

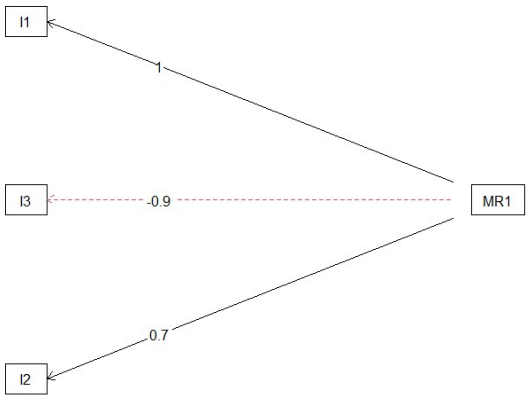
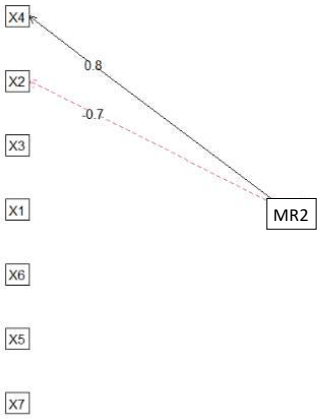


Figure 9

Factor Loading (Manufacturing)



Once these factors were generated, we created a linear regression model to test whether these two factors could predict the dependent variable. When modeling these two factors we did not follow the formerly described methodology of testing the independent variables with different selection and ranking methods. This was because there were only two factors present, meaning that the LASSO results could potentially be biased due lack of independent variables (Heinze, Wallisch, & Dunkler, 2018), the MLR and random forest ranking methods would choose both since there is only one manufacturing and one inventory factor, and backward elimination would return the same results as the linear regression model unless one of the factors were to be found insignificant in predicting the dependent variable. The results of the linear regression model are listed in *Table 8*.

Table 8

Manufacturing & Inventory Factors

Variable	Coefficient	t-value	p-value	VIF
Intercept	0.77273	9.599	2e^-16	
MR1	-0.09289	-3.749	4.19e^-05	1.063121
MR2	0.11445	-6.716	2.10e^-05	1.063121
Median Residual		-0.02028		
R2		0.3774		
Adjusted R2		0.3627		

This model showed a low adjusted R-Squared and a k-fold cross-validation MAE of 0.155. Both metrics which test the fit and accuracy of the model are significantly worse than those of the other models that our team had built previously. For example, the linear regression model using only the variables X_1, X_2, and X_3 had a cross-validated MAE ~27% lower than the model using these two factors. Our team also attempted to split the manufacturing metrics into two and then three factors to be modeled with the one inventory factor but found similar results of low adjusted R-Squared and low k-fold cross-validation MAE for these factor models. From these findings, we concluded that the variability present in the 7-manufacturing metrics is necessary to create a model that produces accurate predictions of the dependent variable.

Understanding that the multicollinearity between the inventory metrics could be resolved using a single inventory factor (MR1), we decided to build a new model using a combination of the data sets' independent variables and the previously generated factors. Using the manufacturing metrics (X_1, X_2… X_7) and the newly generated factor representing the inventory metrics (MR1), we reconducted our four different modeling techniques described earlier to find the recommended subset of KPI metrics to predict our dependent variable (refer to *Table 9*). These subsets were then used in four separate linear regression models, and the repeated k-fold cross-validation was applied to each. The results of the cross-validation using the inventory factor can be found in *Figure 10*, along with the performance of all the other models constructed in our analysis.

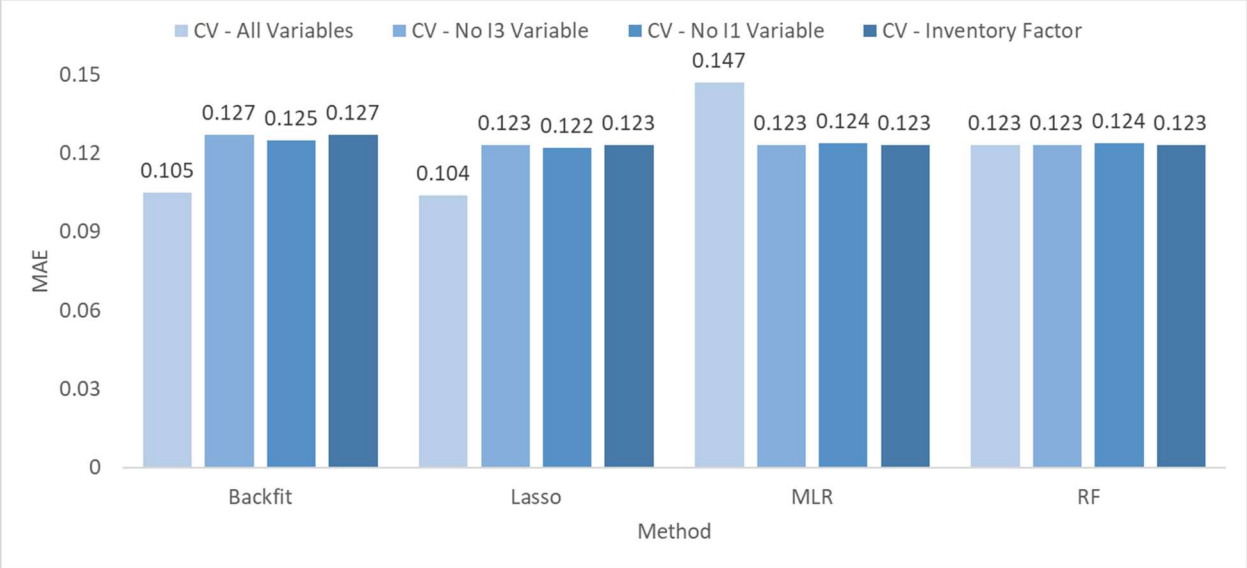
Table 9

Model Results Using Inventory Factor

Model	KPIs Selected
Backward Elimination	X1, X2, X3, X4, X5, MR1
LASSO	X1, X2, X3, MR1
MLR	X1, X2, X3, MR1
Random Forest	X1, X2, X3, MR1

Figure 10

Cross Validated Results of Models Including Inventory Factor



As evident in *Figure 10*, the models with the inventory factor performed within .002 MAE of all models not affected by multicollinearity. Of the four models used to evaluate the dataset that included the inventory factor, three selected variables X_1, X_2, and X_3 as the only manufacturing metrics. The one model that did not (backward elimination) included the variables X_4 and X_5; however, this model was found to be less accurate than the others when tested using the k-fold cross-validation method. With multicollinearity removed from our models, we can proceed with selecting the optimal KPI subset for predicting the dataset's dependent variable in *4.3. KPI Selection*.

4.3. KPI Selection

Considering all available information, we will proceed with the independent variables X_1, X_2, X_3, and I_1 for our final model and sensitivity analysis. Through our modeling analysis, we applied our four different statistical models to four different variations of *Company XYZ's* dataset (all variables, excluding variable I_1, excluding variable I_3, and using an inventory factor), generating sixteen recommended subsets of KPIs for predicting the dependent variable – with the most common selection being X_1, X_2, X_3, and I_1 (25% of all model outputs). The results of the model are provided below in *Table 10.*

Table 10

Final Model Results

Variable	Coefficient	t-value	p-value	VIF
Intercept	0.77273	9.599	4.15e^-15	
X1	-0.2943	-3.749	0.000327	1.008066
X2	-0.50837	-6.716	2.17e^-9	1.036322
X3	0.37126	4.859	5.51e^-6	1.688449
I1	-0.10201	-1.196	0.235192	1.677541
Median Residual		-0.00394		
R2		0.6243		
Adjusted R2		0.6062		

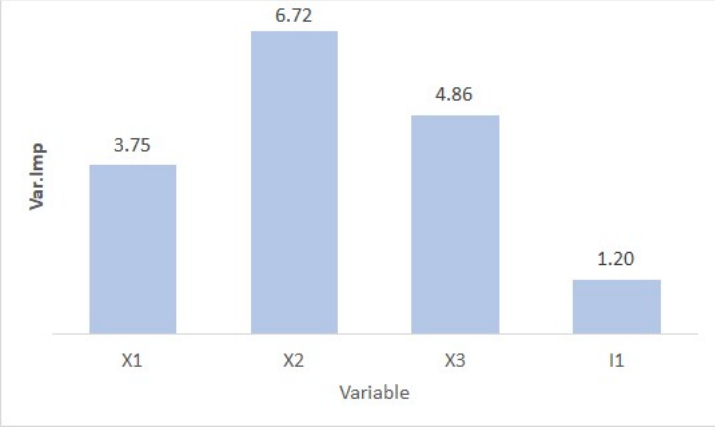
First, we identified that there is no multicollinearity present within this subset of KPIs. As evident in *Table 10*, the VIF score of the four variables falls below ten, meaning that they are likely affected by multicollinearity (O'Brien, 2007). With an adjusted R-squared of 0.6062, greater than 60% of the observed variation in the dependent variable is explained by these four KPIs. The multiple linear regression model is also accurate, with a k-fold cross validation MAE of 0.123. For reference, this model had the second lowest MAE we observed where multicollinearity was not present, and within 0.001 MAE of the best performing model (refer to *Figure 7*).

There are also non-numeric reasons that we selected these metrics for our final model and sensitivity analysis. First, this subset of KPIs is concise – utilizing only four of the original ten variables. This allows for management to focus their attention on the metrics that have the largest impact on their operations, while setting aside other metrics that may not align with the firm's overall objectives. For reference, we have included *Figure 11* which ranks the importance of the four different metrics in predicting the dependent variable. These rankings are based on the absolute value of the t-statistic, the same ranking method discussed in *2.3.1. Multiple Linear Regression*. These variable importance rankings

allow management to quickly identify which manufacturing and inventory metrics have the largest impact on their production efficiency metric. Lastly, these metrics are easily interpretable. Given that these metrics were provided in *Company XYZ's* dataset, we can assume that management is already familiar with and tracking all four of these KPIs.

Figure 11

Final Model Variable Importance



4.3.1. Addressing Insignificant Variables

Though there are many favorable quantitative and qualitative aspects for this subset of KPIs and its subsequent multiple linear regression model, there are potential pitfalls. To begin, the I_1 variable did not have a p-value below our predetermined α value of 0.05. Though it is common practice to eliminate variables found not to be statistically significant, we believe there is enough evidence to include this variable without having an adverse effect on our final analysis. According to the American Statistical Association (ASA), "scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold" (Wasserstein & Lazar, 2016, The ASA Statement on p-Values: Context, Process, and Purpose). From the correlation analysis, it was evident that the variable I_1 has a negative correlation with the dependent variable. Intuitively, it also makes sense that a manufacturing line's inventory position would impact its overall efficiency of operations. According to Louis Columbus (2019), four of the top-ten most valuable metrics to manufacturers are inventory related – carrying cost of inventory, fill rate effectiveness as a percent of all orders, inventory turnover, and supplier quality index. Given the importance of tracking inventory performance within manufacturing processes, our team concluded that it was in the best interest of the model and *Company XYZ* to consider this metric within the sensitivity analysis.

4.3.2. Interaction Effect

Our review of interaction effects also suggested that our team include the I_1 variable. Interaction effect is defined as a scenario within regression where "two or more features/variables combined have a significantly larger effect on a feature as compared to the sum of the individual variables alone" (Khot, 2020, Interaction effect in multiple regression). To conduct this analysis, our team multiplied each of the independent variables in our final model to create six different interaction terms. To test each term, they were included one-by-one in a multiple linear regression model with the four original independent variables. We observed the p-value of the interaction term and the adjusted R-squared of the model to find whether the term was statistically significant or if the explanatory power of the model increased when the term was present in the model. The results of the interaction effect analysis can be found in *Table 11*. These results showed that the only interaction term that had a p-value less than 0.05 and an impact greater than 0.01 on the adjusted R-squared was $X_3 * I_1$. According to the hierarchy principal within variable selection, if an interaction term is found to be significant, the coefficients of the two variables that create the term must be included in the model even if the p-values associated with the individual variables are high (Khot, 2020). This analysis gives further quantitative reasoning behind including the I_1 variable in our final model.

Table 11

Interaction Effect Results

Interaction Terms	P-Value	Adjusted R-Squared
X1 * X2	0.312	0.6064
X1 * X3	0.109	0.6137
X1 * I1	0.151	0.6114
X2 * X3	0.149	0.6114
X2 * I1	0.164	0.6108
X3 * I1	0.001	0.6549

4.4. Sensitivity Analysis

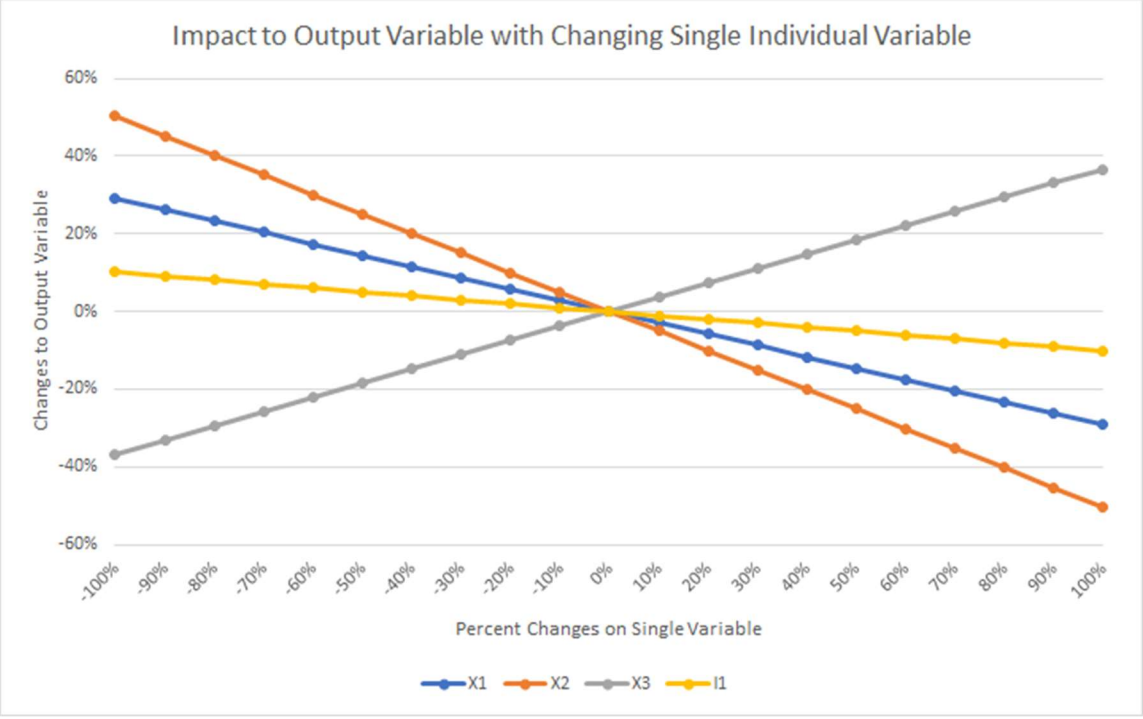
The final model of $y = 0.77273 - 0.2943X1 - 0.50837X2 + 0.37126X3 - 0.10201I1$ provides illustration of the exact relationship between all selected KPIs and the output variable. The model has an intercept of 0.77273, three variables with a negative coefficient, and a positive coefficient for X_3. If *Company XYZ* knows the value of these KPIs, they can estimate the target variable value by plugging the number into the model equation and solve for y.

With the model, it is obvious that I_1 would provide the smallest impact to the target output to compare to other variables. This shows that increasing X_1, X_2, X_3, and I_1 value by 10%, it results to roughly -3%, -5%, 4%, and -1% respectively changes in the output variable. As X_2 has the highest absolute coefficient value, changing the performance of this KPI will provide the most dynamic impact to the output variable, followed by X_3 and X_1.

By changing the single input variable (X_1, X_2, X_3, or I_1), the output variable would align with the coefficient of that individual variable. *Figure 12* illustrates that increasing X_3 performance would positively increase the output variable where the other three variables have the opposite impact due to the opposite coefficient sign. This means that if the goal is to achieve a high positive number on the output variable, it is best to have a high positive X_3 variable, and with X_1, X_2, and I_1 as low as possible. Note that this is a KPI measurement, therefore all variables, X_1, X_2, X_3, and I_1 are expected to be non-negative numbers.

Figure 12

Impact of Output Variable with Changing Single Individual Variable



As the model illustrates, *Company XYZ* can use coefficients to check on the impact of certain variables and identify areas for improvement. As shown in *Figure 11*, the firm can easily see how the target KPI will perform due to a change in a single variable. Also, when *Company XYZ* knows the performance of

the independent variables, the targeted variable value can be calculated using the model. The firm can further elaborate on the sensitivity analysis and develop different scenarios to become more resilient and maintain their target goal. For example, if there is an inventory issue impacting the KPI performance that lowers I_1 by 10%, *Company XYZ* can set up optimization and simulation models to solve for the necessary KPI changes on other variables to maintain at the same output variable target. This simulation can provide insights on how to take action to preserve the output KPI. In this case, the optimization and simulation model can be setup with constraints if one known variable is expected to perform better or worse. For example, we looked at a decrease in performance of I_1, and kept all other independent variables at average levels of performance. Using a simple optimization model, we were able to determine the percentage improvement of all independent variables except I_1 required to maintain a target level of the dependent variable.

Optimization and simulation models can help *Company XYZ* replicate different scenarios and then, based on the context of the KPI, take necessary action to maintain and move in the right direction. This model will be extremely powerful for the company to improve its performance metrics. Similarly, when known changes to these metrics are known in advance, the firm can develop a potential mitigation plan to ensure the best performance.

4.5. Results Summary

We assessed numerous combinations of the performance metrics provided within the dataset to build a multiple linear regression model that could predict a production efficiency metric. Our results showed that the best independent variables to predict this production efficiency metric were variables X_1, X_2, X_3, and I_1. This subset of KPIs were used in a multiple linear regression model to predict the dependent variable, resulting in an adjusted R-squared of 0.6062 and a k-fold cross validation MAE of 0.123. These results tell us that greater than 60% of the observed variation in the dependent variable can be explained using only these four metrics, and the resulting prediction will on average be within 0.123 of the actual value. It also allows the firm to focus on just four metrics to manage their business and achieve their efficiency targets instead of the original ten provided in the dataset.

Utilizing this multiple linear regression model, our team was able to conduct sensitivity analysis with the defined model coefficients. Using a scenario where the performance of the inventory metric I_1 was expected to depreciate by 10%, an optimization model could be made that quickly identifies the targets the firm would need to achieve for variables X_1, X_2, and X_3 to attain their target performance.

This displays a simple yet data-driven approach to help *Company XYZ* set performance standards to meet their overall objectives.

5. DISCUSSION

The central question of this research was to identify how firms can evaluate their current KPIs within their operations to understand which metrics have the greatest impact on performance. Additionally, we looked to determine how these performance metrics could provide insights into forward looking objectives. Assessing the dataset provided by *Company XYZ*, we have been able to clearly answer both questions.

In this chapter, we will discuss the overall significance of our analysis, and how future professionals within the field can apply our team's methodology to assess metrics and provide insight across different organizations and industries. Beginning with managerial insights, we will cover the advantages of using predictive analytics to assess operations metrics and the insights that this methodology provides. Next, we discuss in detail how this framework can be applied outside the context of *Company XYZ*, and the potential improvements that can be made if this analysis were to be replicated.

5.1. Managerial Insights

From our extensive review of literature, we found that it is commonplace for companies to assess the metrics and KPIs that they use to manage their business. Though the methods used for these assessments vary, we found that almost all the various techniques utilized a combination of quantitative and qualitative analysis to find the metrics which provide the most value to the firm. Our proposed methodology of utilizing predictive analytics was a more uncommon approach than other methods such as multi-criteria decision making and linear programming. However, we believe that this method provides a similar level of insight and is more easily replicated within current business environments.

Corporations around the world are making growing investments in big data and business analytics as they seek to make more data-driven decisions across their enterprise. According to Everest Group, the global market for data and data analytics will reach $135 billion by the year 2025 (Bendor-Samuel, 2019). With companies spending vast amounts of capital to manage and derive insights from their data, the field of data science has become a growing area of interest within industry as well. By turning data into business insights, data scientists add immense value to corporations by empowering better decision making,

identifying underlying trends and opportunities, and quantifying business solutions (Monnappa, 2021). Given the growing interest in data science and business analytics, our team found significance in applying some of the tools commonly used by data scientists such as predictive analytics to assess operational metrics within industry.

There are also practical reasons for implementing data science to assess operations metrics. To begin, predictive analytics and statistical modeling are commonly used methods within academia and business. Unlike some of the previously mentioned methods such as multi-criteria decision making, predictive analytics models such as multiple linear regression are commonplace within the curriculums of many MBA and graduate level programs (Murray, 2018). These methods being widely taught means that it is not only likely that someone within an organization possesses the skills to carry out this analysis, but also that many leaders within a company are also familiar with these practices. In contrast, multi-criteria decision-making methods such as Analytical Hierarchy Process (AHP) are not as common in university curriculums, and a firm or organization would be more likely to need an outside group or consultancy to carry out the analysis. Programming languages to execute these predictive models such as R and Python are also free and available to anyone with a computer and internet connection; making them a more accessible and cost-effective solution than some custom software that may be needed for extensive linear programming models such as the one used by Stricker, Minguillon, and Lanza (2017). For these reasons, the only barrier to entry for utilizing this method is the necessary historical data for the predictive models which tracks an organizations performance metrics over time.

Applying our methodology of assessing operations metrics with predictive analytics is not only cost-effective and easily attainable, but it also provides valuable management insights. Today, it is common for various divisions within a company to use different performance metrics to manage their workforce and tasks. For example, leaders in a firm's manufacturing division are likely to use metrics like item fill rates, order cycle times, and carrying costs to gauge the health of their business (Columbus, 2019). On the other hand, chief executives within the same firm are likely to be focused on topline financial metrics such as gross profit, net income, and free cashflows (Key Predictive Indicators, 2010). With two separate sets of KPIs for two different business units, it can be difficult to quantify how a performance metric such as item fill rate is impacting the firm's gross profit. Using the example of *Company XYZ*, we assessed the difference between inventory metrics and manufacturing metrics. Though these two different types of performance metrics are correlated as manufacturing processes are often reliant on the inventory required to produce products, they are managed by two different divisions of the business –

manufacturing and supply chain management. By utilizing predictive analytics, we were able to understand how these two different categories of performance metrics interact to influence the firm's overall objective. Once we identified the metrics that were significant in predicting the dependent variable, we used the coefficients assigned to these metrics through multiple linear programming to quantify the impact these KPIs have on the firm's objective. For example, we found that for *Company XYZ*, a 10% increase in the performance metric X_3 would result in a 4% increase in the firm's production efficiency metric which they aim to maximize.

Using predictive analytics for the assessment of operations metrics also allows firms to identify the metrics which are most important to a firm's overall goals. This analysis can be extremely valuable to a company, as it can identify the metrics that it needs to focus on, eliminate KPIs which are not driving performance, and highlight potential improvements to the already used metrics. It also allows companies to identify which of their key performance indicators are key predictive indicators. According to McKinney Rogers Ltd, key predictive indicators help management teams identify when they are likely to miss their strategic goals before they happen, instead of the traditional key performance indicators which are often focused on past performance (Key Predictive Indicators, 2010). Understanding the metrics that are significant in driving variance in a company's numeric objectives, the firm can begin to act proactively when they notice that one of their key predictive indicators is increasing or decreasing. Utilizing our sensitivity analysis, we were able to identify how changes in the performance metrics X_1, X_2, X_3, and I_1 impact *Company XYZ's* production efficiency metric. Therefore, utilizing these four metrics as key predictive indicators, the firm can begin to predict when their production efficiency will increase or decrease as these four performance metrics change over time.

In conclusion, our methodology and results have provided powerful managerial insights. It is evident that leveraging predictive analytics to assess operations metrics is easily replicated, cost effective, and capable of helping firms prepare for future scenarios. Within a fast-changing business environment, the capability to easily adjust and replicate analysis is essential. For these reasons, we believe that our results have provided valuable insight into *Company XYZ's* dataset, and our methodology will allow the firm to move forward with a clear approach to assess their operational metrics.

5.2. Furthering the Analysis

Through our research, we have set forth a robust methodology for assessing operations metrics through predictive analytics. Though this analysis has been tailored to the dataset provided by *Company XYZ*, our team hypothesizes that this methodology and modeling framework can be used by professionals

and academics across various industries. In this section, we will discuss how similar analysis can be applied to similar problem statements and identify the limitations of our research that could be enhanced by others moving forward.

5.2.1. Replicating the Analysis

When beginning a predictive analytics model, the first two steps are defining business objectives and preparing data (Bari, Chaouchi, & Jung, 2017). Though this may seem straight-forward, these two processes are the rudimentary building blocks for applying our team's methodology. First, an individual must understand what the researched organization's primary target is and how they can quantify it. For *Company XYZ,* their business objective was quantified as an efficiency metric for a given product. However, this objective will likely be different for other companies depending on their value-proposition, business life cycle, and overall corporate objectives. For example, if a new start-up company is focused on expanding sales, their company objective would likely be revenue growth – which is commonly measured monthly, quarterly, or annually (Cross & Wyman, 2011).

Once a numeric target is identified, the next step is to collect data that shows the historical trends of this given metric within the organization. Using the example of revenue growth rate, a firm looking to carry out this analysis would want to collect all available historical data for this metric. This target will become the model's dependent variable. Next, the firm needs to identify all potential metrics or associated datapoints which it believes are impacting their numeric objective. These attributes will become the model's independent variables. Like the objective value, these metrics or datapoints need to be aggregated at appropriate levels that align with the dependent variable. For example, using monthly revenue growth rate as the objective, the other metrics must be aggregated at the monthly level to correctly assess the variance between these data points. *Company XYZ's* dataset aggregated all metrics at a product level; however aggregating performance over time will likely be a more common approach – as measuring performance over various time horizons is common within various industries (Likierman, 2009).

With a dataset that reflects the historic performance of the firm's numeric objective (dependent variable) and related KPIs and metrics (independent variables), the modeling process may begin. Our methodology suggests that the firm assess these independent variables in relation to their dependent variable in two ways: first through feature selection methods to eliminate metrics that are not significant in predicting the dependent variable, and second through ranking the variables' predictive importance to find which metrics are most influential within different categories. By applying these methods, the analyst or researcher will obtain different recommended subsets of metrics to predict the dependent variable. In

this capstone, we selected the machine learning models for selecting and ranking variables that best aligned with the dataset provided by *Company XYZ*.

Once an individual has collected the data, he or she will need to conduct exploratory data analysis to better understand the variables within the dataset, and then select the appropriate algorithms to predict their desired outcomes. For selecting feature selection models like backward elimination and LASSO, any professional looking to replicate this methodology should evaluate different types of wrapper and filter method algorithms to choose the one which is best for the given dataset. As a point of reference, the book *Applied Predictive Methods* (2013) contains a comprehensive review of feature selection algorithms. For selecting an algorithm that will provide the best ranking of variable importance like multiple linear regression and random forest, similar steps should be taken in finding an algorithm that is best fit to a given dataset. Wei, Lu and Song (2015) published a comprehensive review of variable importance analysis which could be used as an initial point of reference for anyone looking to find an appropriate variable ranking model. In summary, regardless of the algorithms one chooses to deploy for the analysis, it is critical that he or she select those that align with their given dataset and problem statement.

With a defined problem statement and selected algorithms and models to rank and select KPIs, the next step is interpreting results and finding the optimal set of KPIs for sensitivity analysis. When selecting this set of metrics, the guiding set of principles should be that the metrics are accurate in predicting the dependent variable, concise, and easily interpretable. By adhering to these three principles, the analysis should provide a subset of KPIs that only include relevant metrics which are statistically proven to impact the firm's objective, with a quantitative measure of how much impact each KPI has on this objective. These results will provide the building blocks for sensitivity analysis, allowing management to understand the impact that increases or decreases in each metric will have on the organization's overall objectives.

5.2.2. Potential Improvements

Throughout our research, we identified a few shortcomings of our analysis that could be improved if this methodology were to be applied elsewhere. First, the dataset provided by *Company XYZ* did not provide any insights into the dates associated with the products. Though we were able to conduct our analysis without this data, we believe that having dates could have potentially enriched our models with external datasets. For example, if dates had been provided, our team could have added independent variables such as stock prices, financial indexes, or unemployment metrics to find if any external macro-

economic adjustments were affecting the firm's objective. Dates could have also provided the context necessary to conduct time-series modeling and identify potential seasonality within the manufacturing processes.

In the case of a financial downturn, natural disaster, or a global pandemic, it is plausible that these external events could impact a firm's manufacturing processes. If our team had access to this data, we would have included the price points of the Dow Jones Industrial Average (DJIA) that corresponded with the production dates of the products as an independent variable. DJIA is one of the most widely viewed benchmarks for economic health, and if this independent variable were to be found significant, it would indicate that the firm's manufacturing processes are impacted by general shifts in the economic landscape. Corresponding stock prices of publicly traded competing firms could also be used to find if there are industry specific trends that could be impacting production. These are just some of the potential external datapoints that could be used to find underlying trends between operations metrics and macro-economic drivers.

Given that our team was assessing inventory metrics within our provided dataset, we would also review price indexes for precious metals and raw materials to find if there is any correlation between these price points and the firm's inventory management. No matter the product or industry, all goods that are manufactured require some form of raw material to produce them – and these raw materials can fluctuate significantly in value. For example, during the COVID-19 global pandemic in April 2020, the benchmark price for hot-rolled coil (HRC) steel, which is commonly used in the automotive industry dropped below $500 per short ton – a price decrease of more than 45% from July 2018 (Ellis, Sierawski, & Trentacosta, 2020). When these fluctuations in raw materials occur, the suppliers who produce and sell these goods often curb production to mitigate financial losses. This causes downstream effects for the manufacturers who procure these goods for their production processes, as lowered production rates of the suppliers can cause shortages in the market. For these reasons, adding the purchasing price of the most commonly used raw materials within a firm's production processes can provide valuable insights into how raw material cost can affect the firm's inventory management.

In conclusion, datapoints that represent time and datasets providing context to external drivers have the potential to enhance predictive modeling. Understanding how macro-economic changes affect a firm's operations can provide additional managerial insights, and help firms create a more robust sensitivity analysis to better plan for future endeavors. For these reasons, we strongly recommend that

Company XYZ (and anyone looking to replicate this methodology) incorporate these datapoints into their modeling.

5.3. Discussion Summary

Our team has identified a comprehensive approach to evaluating KPIs and performance metrics using predictive analytics. We found that machine learning and data science is a fast-growing skill set within the current business landscape, and utilizing techniques within this field to assess operations metrics is attractive for numerous reasons. Our research was able to provide valuable management insights into the how KPIs from different business units interact and quantify their effect on a firm's overall business objective. This analysis also developed a framework for how these metrics can be used proactively to shape business strategies to help firms identify when they could potentially miss their targeted objectives. These results prove that our methodology can be replicated across various industries and can continue to be improved upon with additional datapoints.

6. CONCLUSION

Companies are increasingly pursuing greater analytical capabilities, aiming to drive competitive advantage through customer insights and improved operations. Firms across the world are investing heavily in analytics to draw insights from past performance, predict future performance, and reduce internal complexity. As said by Kiron and Schrage (2019), leaders in today's top firms rely on data to define, communicate, and drive their strategy. Our research focused on the use of key performance indicators (KPIs) in operational management, and how statistical modeling and predictive analytics can be utilized to help identify the best metrics for a firm to gauge performance today and better prepare for the future.

We hypothesized that statistical modeling could help *Company XYZ* distinguish which metrics are most relevant in achieving a numerical objective and could help provide insight into the company's future performance. We used four techniques to select and/or rank KPIs importance – Backward Elimination, LASSO, Multiple Linear Regression, and Random Forest. Of these models, LASSO performed the best with the lowest mean absolute error (MAE) – 0.104. However, given our problem statement, to down-select the few most important KPIs that impact the outcome variable, as well as the issue of multicollinearity of the inventory KPIs, we ended up using a Random Forest model. The Random Forest model MAE is within 0.001 when conducting an analysis to remove correlated variables.

With the variables selected from Random Forest – X_1, X_2, X_3, and I_1, we developed a regression model with adjusted R-squared of 0.61 by eliminating 6 out of 10 variables. With this model, *Company XYZ* will be able to understand how each variable interacts with others by performing sensitivity analysis. Furthermore, they can optimize and simulate different scenarios by adjusting variable inputs to understand whether certain combinations of the variable coefficients are a feasible result or not.

Our methodology, utilizing predictive analytics, was a more uncommon approach than other methods but the result provides a similar level of insight and is more easily replicated within current business environments. With wider usage of the multiple linear regression model, firms can more easily understand the logic behind and can develop plans from the model outcome. Our methodology and model aligned to our hypothesis that statistical modeling can help determine the most important KPIs and help guide strategic actions.

Using predictive analytics to assess operations metrics provides valuable insights to a company as it can identify the metrics that it needs to focus on, eliminate KPIs which are not driving performance, and highlight potential improvements to the already used metrics. This approach also allows companies to identify the key predictive indicators that are significant in driving the company's performance. As time progresses, companies applying this methodology will be able to re-run these modeling practices to assess how their metrics impact operations changes over time. With more data and iterations of the model, the predictive model will likely become stronger and more insightful for the firm. Additionally, anyone replicating this analysis can make potential improvements and strengthen their models by adding datapoints that provide context to external drivers.

Though this research provided valuable insight into the dataset provided by *Company XYZ*, its true power lies outside of this capstone's context. With most large organizations having access to historical records of their performance metrics, and a workforce trained and capable in statistical modeling techniques, this methodology can be applied by numerous organizations and business functions. With a quick analysis of historical performance trends, companies can identify the metrics and macro-economic trends which are truly driving the performance of their operational goals. They can also identify key predictive indicators, which can help them identify potential downturns in performance as certain situations occur. Further, key learning indicators can help a firm measure how these new initiatives are improving performance over time. These techniques will allow management teams to consolidate the metrics that they are currently using and identify potential new KPIs for the firm to leverage. In summary,

predictive analytics provide a fast, cost-effective, and data-driven alternative way of assessing performance metrics.

REFERENCES

Abdallah, T., Diabat, A., & Simchi-Levi, D. (2012). Sustainable supply chain design: A closed-loop formulation and sensitivity analysis. *Production Planning & Control*, *23*(2–3), 120–133. https://doi.org/10.1080/09537287.2011.591622

Akyuz, G. A., & Erkan, T. E. (2010). Supply chain performance measurement: A literature review. INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH, 48(17), 5137–5155. https://doi.org/10.1080/00207540903089536

Barnhart, R. (2012). The Truth & Lies Behind Fill Rates. Tire Review. https://www.tirereview.com/the-truth-lies-behind-fill-rates/

Bendor-Samuel, P. (2019). *Data Analytics and Data Management Market*. Forbes. https://www.forbes.com/sites/peterbendorsamuel/2019/11/26/data-analytics-and-data-management-market/?sh=6fa3cef07678

Brownlee, J. (2016). Tune Machine Learning Algorithms in R (Random Forest Case Study). Machine Learning Mastery. https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/

Brownlee, J. (2018). A Gentle Introduction to k-fold Cross-Validation. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/

Columbus, L. (2019). The 10 Most Valuable Metrics In Smart Manufacturing. Forbes.com. https://www.forbes.com/sites/louiscolumbus/2019/11/20/the-10-most-valuable-metrics-in-smart-manufacturing/?sh=a616d1e2e054

Cross, P., & Wyman, D. (2011). The relationship between monthly, quarterly, and annual growth rates. *Canadian Economic Observer*, *24*(6).

Effendi, M., Khairani, A., & Adnan, R. (2019). Exploratory Factor Analysis (EFA) for Adversity Quotient (AQ) Instrument among Youth. Journal of Critical Reviews. http://www.jcreview.com/fulltext/197-1578466813.pdf

Ellis, N., Sierawski, M., & Trentacosta, J. (2020). *As Raw Material Prices Are Set To Soar, the Effects of 2020 Will Be Felt Into the New Year*. Foley & Lardner LLP. https://www.foley.com/en/insights/publications/2020/12/raw-material-prices-set-soar-effects-of-2020-felt#:~:text=As%20we%20have%20seen%20in,costs%20up%20the%20supply%20chain.

Flovik, V. (2019, June 23). *Artificial Intelligence in Supply Chain Management*. https://towardsdatascience.com/artificial-intelligence-in-supply-chain-management-predictive-analytics-for-demand-forecasting-80d2d512f155.

Fonti, V. (2017). Feature Selection using Lasso. *VU Amsterdam*. https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf

Ford, C. (2016). Getting Started with Factor Analysis. University of Virginia Library. https://data.library.virginia.edu/getting-started-with-factor-analysis/

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician, 63(November 2009), 308–319.

Graham, I., Goodall, P., Peng, Y., Palmer, C., West, A., Conway, P., Mascolo, J. E., & Dettmer, F. U. (2015). Performance measurement and KPIs for remanufacturing. *Journal of Remanufacturing*, *5*(1), 10. https://doi.org/10.1186/s13243-015-0019-2

Hastie, T., & Qian, J. (2014). Glmnet Vignette. Stanford. https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html#:~:text=%23%23%20%5B1%5D%200.08307-,lambda.,standard%20error%20of%20the%20minimum.

Hayes, A. (2020). Multicollinearity. Investopedia. https://www.investopedia.com/terms/m/multicollinearity.asp

Hayes, A., Pakornrat, W., & Khim, J. (2020). Linear Programming. *Brilliant.Org*.

Heinze, G., Wallisch, C., & Dunkler, D. (2018, May). *Variable selection – A review and recommendations for the practicing statistician*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5969114/

Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (20016). The Age of Analytics: Competing in a Data-Driven World. Mckinsey Global Institute. https://doi.org/https://www.mckinsey.com/~/media/McKinsey/Industries/Public%20and%20Social%20Sector/Our%20Insights/The%20age%20of%20analytics%20Competing%20in%20a%20data%20driven%20world/MGI-The-Age-of-Analytics-Full-report.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R (7th ed.). Springer.

Kassambara, A. (2018). Cross-Validation Essentials in R. STHDA. http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/#k-fold-cross-validation

Kenton, W. (2020, September 21). *Multiple Linear Regression (MLR) Definition*. Investopedia. https://www.investopedia.com/terms/m/mlr.asp

Key predictive indicators: The next step for senior management KPIs. (2010). McKinney Rogers. https://mckinneyrogers.blob.core.windows.net/documents-whitepapers/Senior%20Managment%20KPIs.pdf

Kim, B. (2015). Should I always transform my variables to make them normal? University of Virginia Library. https://data.library.virginia.edu/normality-assumption/

Kiron, D. Schrage, M. (2019, June 11). Strategy For and With AI. MIT Sloan Management Review. https://sloanreview.mit.edu/article/strategy-for-and-with-ai/

Kuhfahl, R., Sehlke, C., Sones, J., & Howard, N. (2018). Key Performance Indicators: What Can They Do for You? Armed Forces Comptroller, 63(2), 37–40.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Lakshmanan, S. (2019). How, When, and Why Should You Normalize / Standardize / Rescale Your Data? Towards AI. https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

Likierman, A. (2009). *The Five Traps of Performance Measurement*. Harvard Business Review. https://hbr.org/2009/10/the-five-traps-of-performance-measurement

Liu, F.-H. F., & Liu, Y.-C. (2008). Product line performance assessment on order fulfilment cycle time: A case of microelectronic communication company. *International Journal of Production Research*, *46*(16), 4431–4443. https://doi.org/10.1080/00207540600733543

Mantel, N. (1970). Why Stepdown Procedures in Variable Selection. Technometrics, 12(3), 621–625.

Monnappa, A. (2021). *Why Data Science Matters And How It Powers Business Value*. SimpliLearn. https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article

Munier, N., & Hontoria, E. (2021). General Concepts. In N. Munier & E. Hontoria (Eds.), *Uses and Limitations of the AHP Method: A Non-Mathematical and Rational Analysis* (pp. 1–4). Springer International Publishing. https://doi.org/10.1007/978-3-030-60392-2_1

Murray, S. (2018). How Business Schools Are Teaching Big Data Analytics. BusinessBecause. https://www.businessbecause.com/news/masters-in-business-analytics/5569/how-business-schools-are-teaching-mbas-about-big-data-analytics

Nathan, N. (2019, September 21). *Lasso, ridge and dropout regularization—Their effects on collinearity*. https://towardsdatascience.com/different-forms-of-regularization-and-their-effects-6a714f156521?gi=60a8e4497756

O'Brien, R.,M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality and Quantity, 41(5), 673-690. http://dx.doi.org.libproxy.mit.edu/10.1007/s11135-006-9018-6

Penn State, Eberly College of Science. Applied Multivariate Statistical Analysis (online course), Lesson 12: Factor Analysis. https://online.stat.psu.edu/stat505/lesson/12

Podgórski, D. (2014). Measuring operational performance of OSH management system – A demonstration of AHP-based selection of leading key performance indicators. *Safety Science*, *73*, 146–166.

R Package Documentation (2020). Retrieved October 20, 2020, from https://rdrr.io/

Sensitivity Analysis in Linear Regression (1st ed.). (1988). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316764

Schmidt, K., Aumann, I., Hollander, I., Damm, K., & von der Schulenburg, J.-M. G. (2015). Applying the Analytic Hierarchy Process in healthcare research: A systematic literature review and evaluation of reporting. BMC Medical Informatics and Decision Making, 15(1), 112. https://doi.org/10.1186/s12911-015-0234-7

Sharma, N. (2018). Ways to Detect and Remove the Outliers. Towards Data Science. https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

Urabe, S., Shuang-quan, H., & Munakata, S. (2015). PSI-Cockpit: A Supply-Demand Planning System Focusing on Achieving Goals through Resolving KPI Conflicts.Stefanovic, N. (2014). Proactive

Supply Chain Performance Management with Predictive Analytics. *The Scientific World Journal*, *2014*. https://doi.org/10.1155/2014/528917

Stricker, N., Minguillon, F., & Lanza, G. (2017). Selecting key performance indicators for production with a linear programming approach. *International Journal of Production Research, 55*(19), 5537–5549.

Triantaphyllou, E., Shu, B., Sanchez, S. N., & Ray, T. (1998). Multi-Criteria Decision Making: An Operations Research Approach. *Encyclopedia of Electrical and Electronics Engineering*, *15*, 175–186.

Twin, A. (2020). *Key Performance Indicators (KPIs)*. Investopedia. https://www.investopedia.com/terms/k/kpi.asp

Vandeput, N. (2019). Forecast KPI: RMSE, MAE, MAPE & Bias. Towards Data Science. https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

Wasserstein, R., & Lazar, N. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70(2), 129–133.

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety*, *142*, 399–432.

Wu, S. (2020). Multicollinearity in Regression. Towards Data Science. https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea