Developing a KPI-driven Data Strategy by applying machine learning to KPIs

Vasileiadou Eleftheria

SID: 3308210047

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

April 2023

THESSALONIKI – GREECE

Developing a KPI-driven Data Strategy by applying machine learning to KPIs

Vasileiadou Eleftheria

SID: 3308210047

Supervisor: Dr. Paraskevas Koukaras

Supervising Committee Members: Prof. Christos Tjortjis

Dr. Christos Berberidis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

April 2023

THESSALONIKI – GREECE

Table of Contents

ABSTRACT

Key performance indicators (KPIs) are frequently used by businesses to control their operations. Despite the abundance of metrics, it can be challenging for businesses to pinpoint the ones that are most important to their success. This is an issue for the industry because trivial KPIs can result in management insights that are incorrect.

In today's data-driven world, organizations are constantly seeking ways to improve their performance and stay ahead of the competition. Key Performance Indicators (KPIs) are essential tools that help businesses measure their progress towards achieving their goals. However, with the vast amount of data available, it can be challenging to identify the most relevant KPIs and make informed decisions based on them. This is where machine learning comes in. By applying machine learning algorithms to KPI data, organizations can develop a more data-driven strategy that is tailored to their specific needs. This thesis aims to explore the potential of machine learning in KPI analysis and provide insights into how it can be used to improve business performance.

The strategy involves collecting and analyzing historical data to identify patterns and trends in KPIs, and using this information to develop predictive models that can forecast future KPI performance. This dissertation also discusses the importance of selecting appropriate KPIs as well as the need for ongoing monitoring and refinement of the data strategy. The proposed approach has the potential to improve financial decision-making by providing more accurate and timely insights into KPI performance.

1. INTRODUCTION

Some of the most successful and powerful companies in the current global financial climate have made use of their analytical skills to gain a competitive edge through improved operations and customer insights [1]. As a result of such businesses demonstrating the potential of these cutting-edge technologies, other firms are devoting time and resources to enhancing their skills in areas like analytics, machine learning, and big data.

Finance management has utilized several of these methods to boost financial performance among sectors through improved demand forecasting using machine learning. Although these technologies contribute to wealth creation, they can also make a company's finances more complex. For this reason, in today's highly digitalized world, many business executives look for straightforward, quantifiable metrics to manage their teams. KPIs give businesses the ability to quantify and evaluate the state of the various company functions. Finance professionals are looking into ways to use machine learning to choose the metrics that are ideal for managing a company's finances because they recognize the value of performance metrics and the need for clarity. KPIs can also act as predictive indicators that can support in detecting when a company is going to fall short of their business objectives [2].

While executives and finance specialists rush to apply these strategies, it is crucial to remember that how these practices are used and implemented is just as crucial as the technologies themselves. Money spent on AI and machine learning alone is insufficient. Instead of picking a technology and expecting for a competitive advantage to follow, businesses should concentrate on leveraging these technologies to aid in the development of their strategies and the implementation of these new initiatives. KPIs are particular measurement that businesses should take into account when creating these tools in order to match them with their corporate goals [3].

One of the key benefits of using machine learning to analyze KPIs is that it can help businesses identify patterns and trends that may not be immediately apparent. For example, machine learning algorithms can identify correlations between different KPIs, such as revenue and customer satisfaction, that may not be obvious to human analysts. This can help businesses make more accurate predictions about future performance and identify areas where they need to focus their efforts. Additionally, machine learning can help businesses identify outliers and anomalies in their data, which can be a sign of potential problems or opportunities. Overall, developing a finance KPI data-driven strategy by applying machine learning to KPIs can help businesses make more informed decisions and improve their financial performance.

The use of KPIs is becoming increasingly crucial as technologies advance and management teams have easier access to data. These factors make it clear that companies should look for data-driven methods to identify the appropriate measures to measure their company's success.

Problem Statement

The problem statement highlights the use of a dataset that contains over 200 financial indicators for all the stocks of the US stock market [4]. The dataset was collected from Kaggle and the original financial indicators were scraped from Financial Modeling Prep API by the Kaggle author. The objective of the study is to evaluate the impact of metrics from financial statements, such as profit and loss and balance sheet, on forecasting revenue using machine learning. The analysis and logic used in this study can be applied by any company by adjusting the KPIs according to their needs.

The study aims to help companies identify the KPIs that are driving success and remove those that are not. With access to thousands of data points, companies can gain valuable insights into their finance management and streamline their performance measurements. By narrowing down the list of the most important performance measurements, companies can minimize confusion caused by insignificant measures. This approach can help companies optimize their operations and improve their overall performance.

Overall, the problem statement highlights the importance of using data-driven approaches to evaluate financial indicators and identify the most important KPIs for a company's success. By leveraging machine learning and analyzing financial statements, companies can gain valuable insights into their operations and streamline their performance measurements. This approach can help companies optimize their operations and improve their overall performance, ultimately leading to increased revenue and profitability.

Though there are numerous frameworks and methods for evaluating and ranking KPIs, this study will use machine learning techniques to determine which among the provided measures is the most statistically significant when taking the company's primary objective into account. This strategy will be consistent with conventional predictive modeling methods which involve choosing the outcome parameter the business wants to enhance and then performing analysis to determine which independent features have the greatest influence on this dependent variable. Which of the finance metrics has the greatest influence on the revenue? That question will be answered by this study.

My theory is that machine learning techniques can help the business determine which indicators prove most important in reaching a quantifiable goal and can assist in predicting how it will perform in the future. The results of this study will show how to use different machine learning techniques on an example data set. It will also discuss various statistical models' fundamental underlying methodologies. The explanations and cases offered can be used as a guideline.

2. THEORETICAL BACKROUND

In order to monitor revenue performance, we need to choose the KPIs that will have the greatest effect from a list of existing metrics. The first section of this literature review will define a KPI and explain how it relates to Finance. This will be followed by a short discussion of the methods currently in use for choosing and ranking KPIs. The next section will discuss machine learning-based alternatives to the current ranking and selection procedures. The conclusion of this evaluation will clearly outline the next step that can be taken in order to remedy the stated problem.

2.1. KPIs in Finance

A key performance indicator is a measure of quality that allows success of various businesses or employees to be compared across baselines. KPIs should be connected to profitability or a strategic objective to show that reaching a certain figure is important and capable of illuminating a precise goal for the company to work toward. The most important aspects of KPIs are that they shouldn't be overused or overly complex, as both of these can bring users astray. The net profit margin, that transforms net profit to a revenue percentage, is one widely used KPI [5].

In spite of the development of new tools, KPIs are still frequently used today. When they are chosen and created properly, they serve as a very useful support to the domains that need attention and can identify possible underlying issues [6]. However, KPIs can be changed to better suit business requirements as novel innovations and even more data become available. New methods of data collection make it possible to gather more data and build fresh KPIs for evaluating effectiveness. As an example, adding a video sensor may help gather more information for analysis and help create new KPIs, like more precise root causes of damaged shipments.

However, there can exist drawbacks to relying entirely on KPIs for guidance in decisions and assessing progress toward a goal. Completely focusing on achieving outstanding KPI results might not lead to the right outcome Furthermore, KPIs aren't always consistent with one another. Departments occasionally may have opposing goals in reaching excellent KPI outcomes [7].

Finance professionals are constantly seeking ways to improve their decision-making processes. One of the most effective ways to achieve this is by using Key Performance Indicators (KPIs) to measure and track financial performance. However, with the increasing volume of financial data available, it can be challenging to identify the most relevant KPIs and interpret the data accurately. This is where machine learning comes in. By leveraging advanced algorithms and statistical models, machine learning can help finance professionals identify the most critical KPIs and provide insights into financial performance that would be difficult to obtain through traditional methods. In this article, we will explore the role of machine learning in KPI analysis and how it can help finance professionals make better decisions.

Organizations can choose from a wide variety of KPIs, so the issue is: how can analytical techniques be most effectively used to choose the KPI that will have the biggest impact on the underlying management focus? We'll go over a few of the methods that are frequently used to choose and forecast performance measures in the section that follows.

2.2. Financial Statements in scope

The metrics in this study will be categorized based on their location in either the profit and loss statement or the balance sheet. To accomplish this, it is crucial to understand the significance of these financial statements and the information they provide. Therefore, it is necessary to gain a comprehensive understanding of the purpose and importance of these statements.

Profit and loss statement, often called an income statement, displays a company's sales, costs, and net gain or loss during a given time period. It is a summary of a business's financial performance over a specific time period, typically a quarter or a year. The statement shows the company's revenue, cost of goods sold, gross profit, operating expenses, and net income or loss.

The assets, liabilities, and equity of a firm are displayed on a balance sheet, which is a financial statement. It provides a snapshot of a company's financial position at a particular moment. The statement shows the company's assets, such as cash, accounts receivable, inventory, and property, plant, and equipment. It also shows the company's liabilities, such as accounts payable, loans, and other debts. Finally, it shows the company's equity, which is the difference between its assets and liabilities.

In summary, the profit and loss statement show a company's financial performance over a specific period, while the balance sheet shows its financial position at a particular point in time. Both statements are essential for understanding a company's financial health and making informed decisions about its future.

2.3. Current Methods for KPI Review and Selection

Key Performance Indicators (KPIs) are essential tools for measuring the success of any business or organization. They provide valuable insights into the performance of various aspects of the business, enabling decision-makers to make informed decisions and take appropriate actions. However, selecting the right KPIs can be a daunting task, as there are numerous factors to consider, such as the business objectives, industry standards, and available data sources. To address this challenge, various methods have been developed for KPI review and selection. These methods can be broadly classified into three categories: data-driven approach, expert-driven approach, and hybrid approach. We will explore each of these methods in detail and discuss their advantages and limitations. [8]

2.3.1. Data-driven approach

A data-driven approach is a method of decision-making that relies on analyzing historical data to identify the most relevant key performance indicators (KPIs) for a particular business or industry. This approach involves using machine learning algorithms to identify patterns and correlations in the data, which can help to identify the most important KPIs. [9]

The first step in a data-driven approach is to gather and analyze historical data. This data can come from a variety of sources, including customer data, sales data, financial data, and operational data. Once the data has been collected, it is analyzed using machine learning algorithms to identify patterns and correlations.

The next step is to identify the most relevant KPIs for the business or industry. This can be done by analyzing the data and identifying the metrics that have the greatest impact on the business. For example, if a business is focused on increasing sales, the most relevant KPIs might be revenue, customer acquisition, and customer retention.

Once the most relevant KPIs have been identified, they can be used to guide decision-making and measure performance. By tracking these KPIs over time, businesses can identify trends and make data-driven decisions to improve their performance.

Overall, a data-driven approach is a powerful tool for businesses and industries looking to improve their performance. By analyzing historical data and identifying the most relevant KPIs, businesses can make informed decisions and drive growth and success.

In our study of the literature, we identified two distinct approaches: multi-criteria decision making and linear programming. (MCDM).

2.3.1.1. Multi-Criteria Decision Making

Operations research frequently employs a technique known as multi-criteria decision making (MCDM), also known as multi-criteria decision analysis (MCDA), to assist end users in making the best choice by weighing a variety of factors [10]. MCDM can be divided into two additional groups; multi-attribute decision making (MADM), that concentrates on issues where the set of choice options have been previously established., and multi-objective decision making (MODM), that is utilized when there are several objective functions. Data envelopment analysis and the analytical hierarchy process are two instances of MADMs that have been used to evaluate operations metrics in the literature that we reviewed.

Data envelopment analysis (DEA) measures the input output effectiveness relation on how well an objective item is at the optimal level using a linear programming-based model. DEA is frequently used in manufacturing settings to basically create a frontier, also known as the best trajectory output line. The output is not operating at its best if its performance is not on the frontier line. This can help management determine which quantifiable items need to be improved and in which way [11]. Instead of determining the relative importance of different input variables, this method analyzes and assesses the effectiveness of certain outputs based on a variety of input variables and the extent to which the result might be enhanced.

The DEA technique is effective at assessing efficiency performance and identifying areas that need improvement. However, because the method combines all variables to make a formulation, it is not appropriate for determining the significance of specific parameters or choosing and ranking the input metric that could have the biggest effect on the final KPI. There are also hybrid DEA versions that combine various theories. These theories, however, continue to be grounded in DEA, which is irrelevant to ranking and choosing the essential performance measurements which have the biggest effects on the goal function.

The most popular MCDM technique is called the analytical hierarchy process (AHP), and thousands of papers have been written on its various uses [12]. AHP is a framework for making decisions that enables individuals to divide issues into stages. [13]. According to a set of standards they called SMART (specific, measurable, achievable, relevant, and time-bound), a comparison of all the indicators and KPI groups in pairs was created and found to produce a vector where every component has a weighting. A ranking of the 20 areas' indicators performance emerged [14]. The study demonstrated the usefulness of a rank-based method for choosing KPIs; nevertheless, considering the dataset's limitations, other techniques must be used to rank the different metrics in our study accurately.

2.3.1.2. Linear Programming

Finding the best answer to a goal function provided multiple linear constraints is frequently done using linear programming. [15]. Integer linear programming, wherein only integers may be taken into account for the function's variables of choice, and mixed integer linear programming, wherein just a few variables of choice are integers, are two examples of different types of linear programming. A strategy centered on reducing the number of measures required to satisfy the information needs they has been established. This process was chosen because it uses a mathematical equation to determine how many KPIs to use, as opposed to alternative processes like the analytical hierarchy process, which focuses primarily on subjective metrics. It was recognized that KPI selection is frequently subjective in nature[16]. The following step involved measuring and choosing KPIs according to three factors: sensitivity, explanatory power, and interval divergence. These results were computationally derived, the project's challenge and dataset render this approach inappropriate for the project because it requires a carefully constructed dataset and specialized software.

2.3.2. Expert-driven approach

The expert-driven approach is a method of identifying the most important key performance indicators (KPIs) for a particular business or industry. This approach involves consulting with subject matter experts who have extensive knowledge and experience in the field to determine the most relevant KPIs that should be tracked and measured.

The experts are typically individuals who have a deep understanding of the industry, the business processes, and the key drivers of success. They may include executives, managers, analysts, or consultants who have worked in the industry for many years and have a wealth of knowledge and expertise.

The first step in the expert-driven approach is to identify the experts who will be consulted. These experts may be internal or external to the organization, and they should have a broad range of knowledge and experience in the industry. Once the experts have been identified, they are typically interviewed or surveyed to gather their insights and opinions on the most important KPIs for the business or industry.

The next step is to use machine learning algorithms to validate and refine the KPIs identified by the experts. Machine learning algorithms can analyze large amounts of data and identify patterns and trends that may not be immediately apparent to human experts. This can help to ensure that the KPIs identified are truly the most important ones for the business or industry.

Overall, the expert-driven approach is a valuable method for identifying the most important KPIs for a business or industry. By consulting with subject matter experts and using machine learning algorithms to validate and refine the KPIs, organizations can ensure that they are tracking the metrics that truly matter and are most likely to drive success.

2.3.3. Hybrid approach

The hybrid approach is a method that combines the strengths of both data-driven and expert-driven approaches to identify the most important key performance indicators (KPIs) for a business. The data-driven approach involves using machine learning algorithms to analyze historical data and identify potential KPIs. This approach is useful because it can uncover patterns and relationships in the data that may not be immediately apparent to human experts.

However, the data-driven approach has limitations. It may identify KPIs that are not relevant to the business or that do not align with the business's goals and objectives. This is where the expert-driven approach comes in. Subject matter experts can review the potential KPIs identified by the machine learning algorithms and validate them based on their knowledge and experience. They can also refine the KPIs to ensure that they are relevant and aligned with the business's goals.

The hybrid approach is beneficial because it combines the strengths of both approaches. It allows businesses to leverage the power of machine learning algorithms to analyze large amounts of data and identify potential KPIs, while also ensuring that the KPIs are relevant and aligned with the business's goals. This approach can help businesses make more informed decisions and improve their performance by focusing on the most important KPIs.

2.4. Feature Selection and Ranking in Machine Learning Models

Machine learning has revolutionized the way we approach data analysis and decision-making. One of the key challenges in machine learning is selecting the most relevant features from a large set of potential predictors. Feature selection and ranking are critical steps in the machine learning pipeline, as they help to reduce the dimensionality of the data and improve the accuracy and interpretability of the models. In recent years, there has been a growing interest in developing machine learning models specifically for feature selection and ranking. These models use various techniques such as filter, wrapper, and embedded methods to identify the most informative features and rank them based on their importance. We will explore some of the popular machine learning models for feature selection and ranking and discuss their strengths and limitations.

2.4.1. Backward Elimination

Our study will now focus on techniques for choosing these metrics now that two distinct strategies for ranking KPIs have been identified. A common step function method called forward selection is comparable to backward elimination. In both cases, a collection of independent variables is recursively regressed against the dependent variable until only those independent variables that are statistically significant are left [19].

Backward elimination starts with all the independent variables provided, eliminates the one that is least significant when estimating the dependent variable, updates the model with the available independent variables, and continues this process until there aren't any more insignificant independent variables left [19]. The process of forward selection starts with the construction of a univariable model for each independent variable, followed by selecting the univariable model with the highest level of significance, the regressing of the remaining independent variables, as well as the selection of the most important independent variables to be included to the model until there are not any additional independent variables to include. Although both of these methods can identify important variables, many statisticians favor the backward elimination model when the independent variables are collinear [20].

Being that it functions as an extension for numerous linear regression models, backward elimination is also a compelling feature selection algorithm. MLR has become one of the most widely used statistical modeling methods in use nowadays, as was previously mentioned. Such models give analysts and businesses visibility on the manner that each independent variable affects the prediction of a dependent variable through individually assigned p-values and coefficients. This feature selection method will be easy, affordable, and understandable because backward elimination methods are accessible in many open-source R packages.

2.4.2. Least Angle Selection and Shrinkage Operator

This is also a feature selection method. Although this method could be applied to any dataset with at least ten independent variables, LASSO is typically employed with high-dimensional datasets at which the quantity of independent variables exceeds the quantity of individual records. [19].

Regularization is how LASSO technique chooses features for a model [20]. The factors that have a non-zero coefficient at the conclusion of the shrinkage process are those that are chosen for the model. Regularization is a potent idea that, by reducing collinearity between chosen independent variables, diminishes prediction error and lowers the possibility of model to overfit [22]. Similar to LASSO, ridge regression also uses a shrinkage technique, but only reduces rather than eliminates the regression coefficients [21].

2.4.3. MLR

Multiple linear regression (MLR) is a statistical modeling technique that is frequently employed. MLR is a statistical method that "uses various explanatory variables to forecast the result of a response variable." [17]. The independent variables in this meaning are going be the variables that explain the outcome, and the dependent variable is going be the response variable. MLR generates a forecast for the dependent variable using the values of the independent variables and is frequently employed when the amount of observations is greater than the number of variables [18]. With the help of MLR, users can quickly understand the importance of a given value when predicting the result of the response variable (p-value) as well as the effect that the variables that explain the outcome have on the response variable as it changes in measure. (coefficient). Professionals from a variety of sectors frequently use MLR due to its simple interpretation. This model does have some drawbacks, though. For instance, MLR suggests that the independent variables don't have an overly correlated and that the dependent and independent variables have a linear relationship. Therefore, when using this method, appropriate modeling techniques like testing for linearity and multicollinearity have to be used.

There are several methods to gauge an independent variable's significance within an MLR model. However, there is a straightforward method to determine this measure using software libraries, like the caret package in R. Every parameter is ranked according to the absolute value of the t-statistic using the caret technique. Each independent variable will be given a ranking by this method, which may be compared to the outcomes of the variable rankings supplied from different statistical models.

2.4.4. Random Forest

This model analyzes datasets with a variety of observations and variables and ranking the significance of each variable [18]. With Random Forest models, nonlinearities can be evaluated without having to formally model them, unlike with MLR. As a tree-based model, it is dependent on regression trees that continuously divide the data, resulting in a regression function that is constructed on a multidimensional step function.

There are two methods for ordering Random Forest variables. The Gini importance method, which ranks features based to their average impurity reduction whenever variables are split, was the first variable importance measure examined. Although this method of variable ranking is frequently employed, some split variables may cause the variable importance measures to become biased. The second technique, known as MSE reduction, ranks variables based on how well they forecast outcomes. Using R's Random Forest package, both metrics can be discovered.

2.5. Literature Review Outline

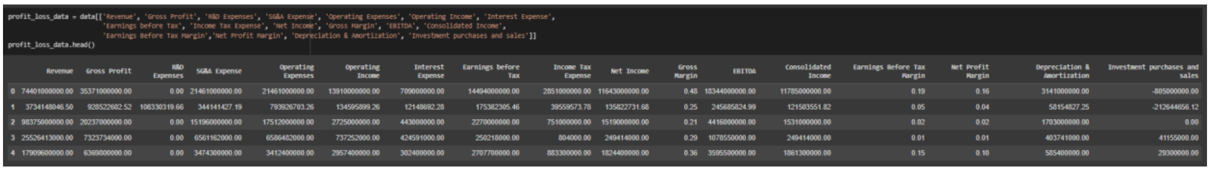
Overall, machine learning can be a powerful tool for KPI review and selection, as it can help to identify patterns and correlations in data that may not be immediately apparent to human analysts. However, it is important to ensure that the machine learning algorithms are properly trained and validated to ensure that the KPIs identified are accurate and relevant.

Our analysis considered backward elimination and LASSO as techniques for choosing and removing KPIs from the dataset, as well as multiple linear regression and random forest as techniques for ranking the KPIs. The focus of this study will now shift to using these machine learning techniques to address the stated issue.

3. IMPLEMENTATION FRAMEWORK

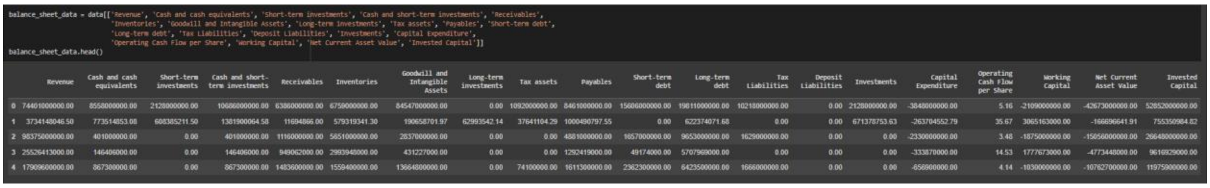
The first part of this study examines the dataset and talks about the procedures used to verify and clean the data. The four various modeling approaches are then thoroughly discussed, and we then talk about how we will compare the results for each model. This section will outline a clear course of action for addressing the issue statement of determining the most effective way to choose KPIs using statistical modeling and machine learning.

The methodology used in this dataset involves collecting financial indicators for all the stocks of the US stock market. The dataset contains over 200 financial indicators that have been scraped from the Financial Modeling Prep API [4]. These indicators are the same ones found in the 10-K filings that publicly traded companies release yearly. The dataset covers the date range of 2014-2018 and contains 22077 observations. Originally, there were 224 columns in the dataset. but the author created two groups of columns, one for the metrics found in the balance sheet and one for the metrics found in the profit and loss statements [Figure 3]. Overall, using financial indicators found in the 10-K filings, the author is able to create a comprehensive dataset that covers a wide range of metrics. By focusing on the metrics that have the biggest impact on revenue, the author is able to create a more accurate forecast. The table included in the dataset makes it easy to understand which metrics are included and where they are found in the financial statements.



```
profit_loss_data = data[['Revenue', 'Gross Profit', 'R&D Expenses', 'SG&A Expense', 'Operating Expenses', 'Operating Income', 'Interest Expense',
                       'Earnings before Tax', 'Income Tax Expense', 'Net Income', 'Gross Margin', 'EBITDA', 'Consolidated Income',
                       'Earnings Before Tax Margin','Net Profit Margin', 'Depreciation & Amortization', 'Investment purchases and sales']]
profit_loss_data.head()
```

	Revenue	Gross Profit	R&D Expenses	SG&A Expense	Operating Expenses	Operating Income	Interest Expense	Earnings before Tax	Income Tax Expense	Net Income	Gross Margin	EBITDA	Consolidated Income	Earnings Before Tax Margin	Net Profit Margin	Depreciation & Amortization	Investment purchases and sales
0	74401000000.00	35371000000.00	0.00	21461000000.00	21461000000.00	13910000000.00	709000000.00	14494000000.00	2851000000.00	11643000000.00	0.48	18344000000.00	11785000000.00	0.19	0.16	3141000000.00	-805000000.00
1	3734148046.50	928522602.52	108330319.66	344141427.19	793926703.26	134595899.26	12148692.28	175382305.46	39559573.78	135822731.68	0.25	245685824.99	121503551.82	0.05	0.04	58154827.25	-212644656.12
2	98375000000.00	20237000000.00	0.00	15196000000.00	17512000000.00	2725000000.00	443000000.00	2270000000.00	751000000.00	1519000000.00	0.21	4416000000.00	1531000000.00	0.02	0.02	1703000000.00	0.00
3	25526413000.00	7323734000.00	0.00	6561162000.00	6586482000.00	737252000.00	424591000.00	250218000.00	804000.00	249414000.00	0.29	1078550000.00	249414000.00	0.01	0.01	403741000.00	41155000.00
4	17909600000.00	6369800000.00	0.00	3474300000.00	3412400000.00	2957400000.00	302400000.00	2707700000.00	883300000.00	1824400000.00	0.36	3595500000.00	1861300000.00	0.15	0.10	585400000.00	29300000.00



```
balance_sheet_data = data[['Revenue', 'Cash and cash equivalents', 'Short-term investments', 'Cash and short-term investments', 'Receivables',
                         'Inventories', 'Goodwill and Intangible Assets', 'Long-term investments', 'Tax assets', 'Payables', 'Short-term debt',
                         'Long-term debt', 'Tax Liabilities', 'Deposit Liabilities', 'Investments', 'Capital Expenditure',
                         'Operating Cash Flow per Share', 'Working Capital', 'Net Current Asset Value', 'Invested Capital']]
balance_sheet_data.head()
```

	Revenue	Cash and cash equivalents	Short-term investments	Cash and short-term investments	Receivables	Inventories	Goodwill and Intangible Assets	Long-term investments	Tax assets	Payables	Short-term debt	Long-term debt	Tax Liabilities	Deposit Liabilities	Investments	Capital Expenditure	Operating Cash Flow per Share	Working Capital	Net Current Asset Value	Invested Capital
0	74401000000.00	8558000000.00	2128000000.00	10686000000.00	6386000000.00	6759000000.00	84547000000.00	0.00	1092000000.00	8461000000.00	15606000000.00	19811000000.00	10218000000.00	0.00	2128000000.00	-3848000000.00	5.16	-2109000000.00	-42673000000.00	52852000000.00
1	3734148046.50	773514853.08	608385211.50	1381900064.58	11694866.00	579319341.30	190658701.97	62993542.14	37641104.29	1000490797.55	0.00	622374071.68	0.00	0.00	671378753.63	-263704552.79	35.67	3065163000.00	-166696641.91	755350984.82
2	98375000000.00	401000000.00	0.00	401000000.00	1116000000.00	5651000000.00	2837000000.00	0.00	0.00	4881000000.00	1657000000.00	9653000000.00	1629000000.00	0.00	0.00	-2330000000.00	3.48	-1875000000.00	-15056000000.00	26648000000.00
3	25526413000.00	146406000.00	0.00	146406000.00	949062000.00	2993948000.00	431227000.00	0.00	0.00	1292419000.00	49174000.00	5707969000.00	0.00	0.00	0.00	-333870000.00	14.53	1777673000.00	-4773448000.00	9616929000.00
4	17909600000.00	867300000.00	0.00	867300000.00	1483600000.00	1559400000.00	13664800000.00	0.00	74100000.00	1611300000.00	2362300000.00	6423500000.00	1666000000.00	0.00	0.00	-656900000.00	4.14	-1030000000.00	-10762700000.00	11975900000.00

Figure 3: Profit & Loss and Balance Sheet Dataframes

3.1. *Preprocessing*

Prior to applying statistical modeling to a dataset, it is crucial to carefully examine the data points for abnormalities and trends which might have an impact on the model's results.

Preprocessing is crucial in machine learning because it helps to improve the quality of the data and make it more suitable for analysis. The main reasons why preprocessing is important are [23]:

1. Data cleaning: Preprocessing helps to remove any irrelevant or duplicate data, as well as any outliers or errors that may affect the accuracy of the model.

2. Data normalization: Preprocessing helps to scale the data so that it is consistent and comparable across different features. This is important because some features may have a larger range of values than others, which can skew the results.

3. Feature selection: Preprocessing helps to identify the most important features that are relevant to the problem being solved. This can help to reduce the dimensionality of the data and improve the accuracy of the model.

4. Data transformation: Preprocessing helps to transform the data into a format that is more suitable for analysis. For example, text data may need to be converted into numerical data before it can be used in a machine learning model.

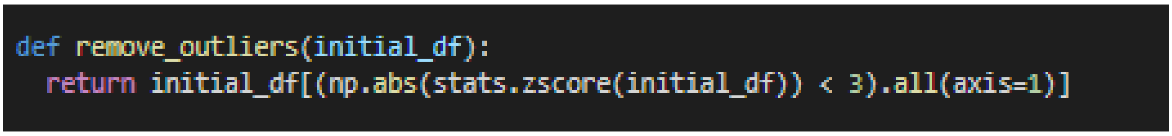
Overall, preprocessing is crucial in machine learning because it helps to ensure that the data is accurate, consistent, and relevant to the problem being solved. This can help to improve the accuracy of the model and make it more effective in making predictions or classifications.

3.1.1. Removing Outliers

The provided dataset had outliers that needed to be removed, so we will briefly discuss the significance of this step. If an entry in a dataset experienced an anomaly, or if the data were incorrectly collected, an outlier may appear.

A ranking above one for a measure that ranges between zero and one, for instance, would likely represent an error in the data gathering, whereas a really bad performance for one specific line-item would possibly be recognized as an outlier.

Using mathematical functions like z-scores or interquartile range (IQR) to find and remove outlier entries from a dataset is a prevalent practice. These mathematical procedures determine the distribution of the data, and they then omit records that are discovered outside of an analyst-specified boundary [24]. The statistical models' capacity to detect variability between variables will be reduced by eliminating these outliers. But it also generates a dataset that is less skewed by anomalies and potential data collection errors, which eventually results in a more accurate predictive model.



```
def remove_outliers(initial_df):
  return initial_df[(np.abs(stats.zscore(initial_df)) < 3).all(axis=1)]
```

Figure 3.1.1: Removing Outliers Function

In this case a function in Python was used that removes outliers from a given DataFrame called "initial_df" [Figure 3.1.1]. The function uses the z-score method to identify outliers. The z-score is a measure of how many standard deviations a data point is from the mean. The function first calculates the z-score for each data point in the DataFrame using the "stats.zscore" function from the "scipy" library. Then, it checks if the absolute value of the z-score is less than 3 for each data point. If the z-score is greater than 3, it is considered an outlier. Finally, the function returns a new DataFrame that only includes rows where all the z-scores are less than 3. This effectively removes any outliers from the original DataFrame.

3.1.2. Scaling Variables

Scaling variables in machine learning refers to the process of transforming the values of input features to a similar scale. This is done to ensure that all features contribute equally to the model and to prevent features with larger values from dominating the model.

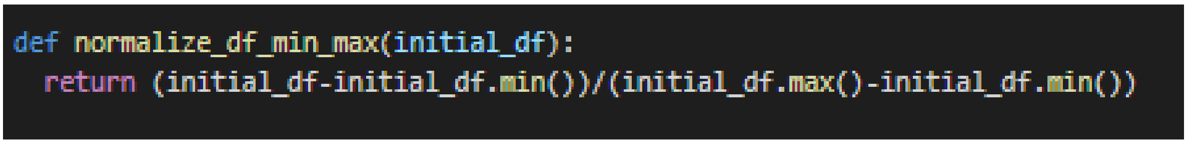
There are different methods for scaling variables, including [25]:

1. Standardization: This involves transforming the values of the features to have a mean of zero and a standard deviation of one. This method is useful when the data is normally distributed.

2. Min-max scaling: This involves transforming the values of the features to a range between 0 and 1. This method is useful when the data is not normally distributed.

3. Robust scaling: This involves transforming the values of the features to a range between -1 and 1, based on the median and interquartile range. This method is useful when the data contains outliers.

Scaling variables is an important preprocessing step in machine learning, as it can improve the performance of the model and prevent issues such as overfitting.



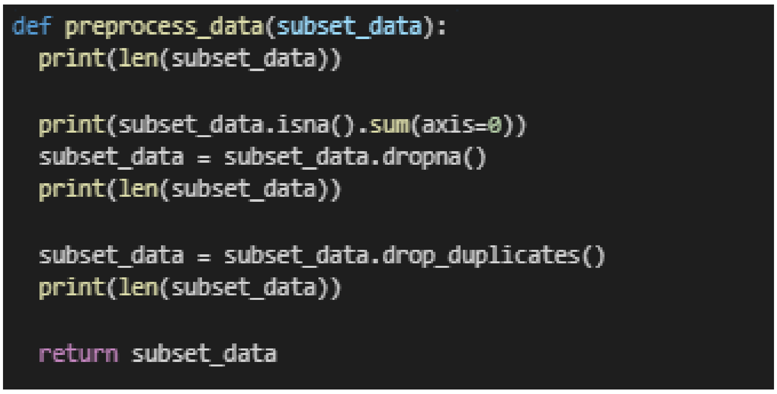
```
def normalize_df_min_max(initial_df):
  return (initial_df-initial_df.min())/(initial_df.max()-initial_df.min())
```

Figure 3.1.2: Normalizing Function

In this case a function in Python was used that takes an initial dataframe as input and returns a normalized dataframe using the min-max scaling technique [Figure 3.1.2]. The function first subtracts the minimum value of each column from the initial dataframe and then divides the result by the range of each column (i.e., the difference between the maximum and minimum values). This ensures that all values in the dataframe are scaled between 0 and 1. The resulting normalized dataframe is then returned by the function.

3.1.3. Removing missing values

Removing null values is important in machine learning because null values can negatively impact the accuracy and reliability of the model. Null values can cause errors in the analysis and prediction process, leading to incorrect results and unreliable insights. Additionally, null values can skew the distribution of data, leading to biased results and inaccurate predictions. By removing null values, the model can be trained on a completer and more accurate dataset, leading to better performance and more reliable predictions. Overall, removing null values is a crucial step in ensuring the accuracy and reliability of machine learning models.



```
def preprocess_data(subset_data):
  print(len(subset_data))

  print(subset_data.isna().sum(axis=0))
  subset_data = subset_data.dropna()
  print(len(subset_data))

  subset_data = subset_data.drop_duplicates()
  print(len(subset_data))

  return subset_data
```

Figure 3.1.3: Removing null values

In this case, a function in Python is used that takes a subset of data as input and performs some preprocessing steps on it [Figure 3.1.3]. The first line of the function defines the function name and the input parameter. Then it uses the isna() function to check for missing values in the subset_data and then sums them up along the columns (axis=0). The result is printed to the console. It drops any rows that contain missing values from the subset_data and prints the new length of the data. Finally, it drops any duplicate rows from the subset_data and prints the new length of the data.

3.2. Exploratory Data Analysis (EDA)

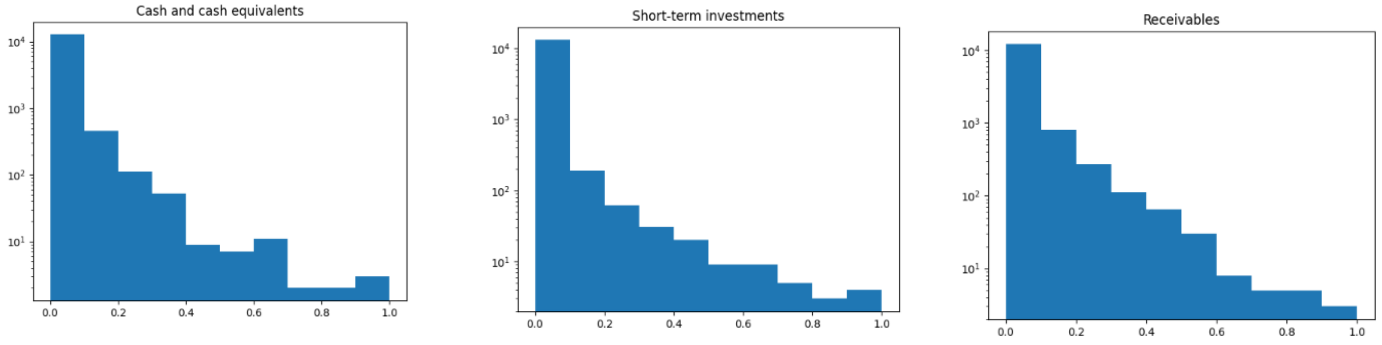
Exploratory data analysis is the term used to describe this pre-modeling procedure. (EDA). This part of the analysis will go over the steps required to run EDA on the specific dataset in order to determine the distribution and correlation of the variables, along with how to get rid of outliers and normalize the data points.

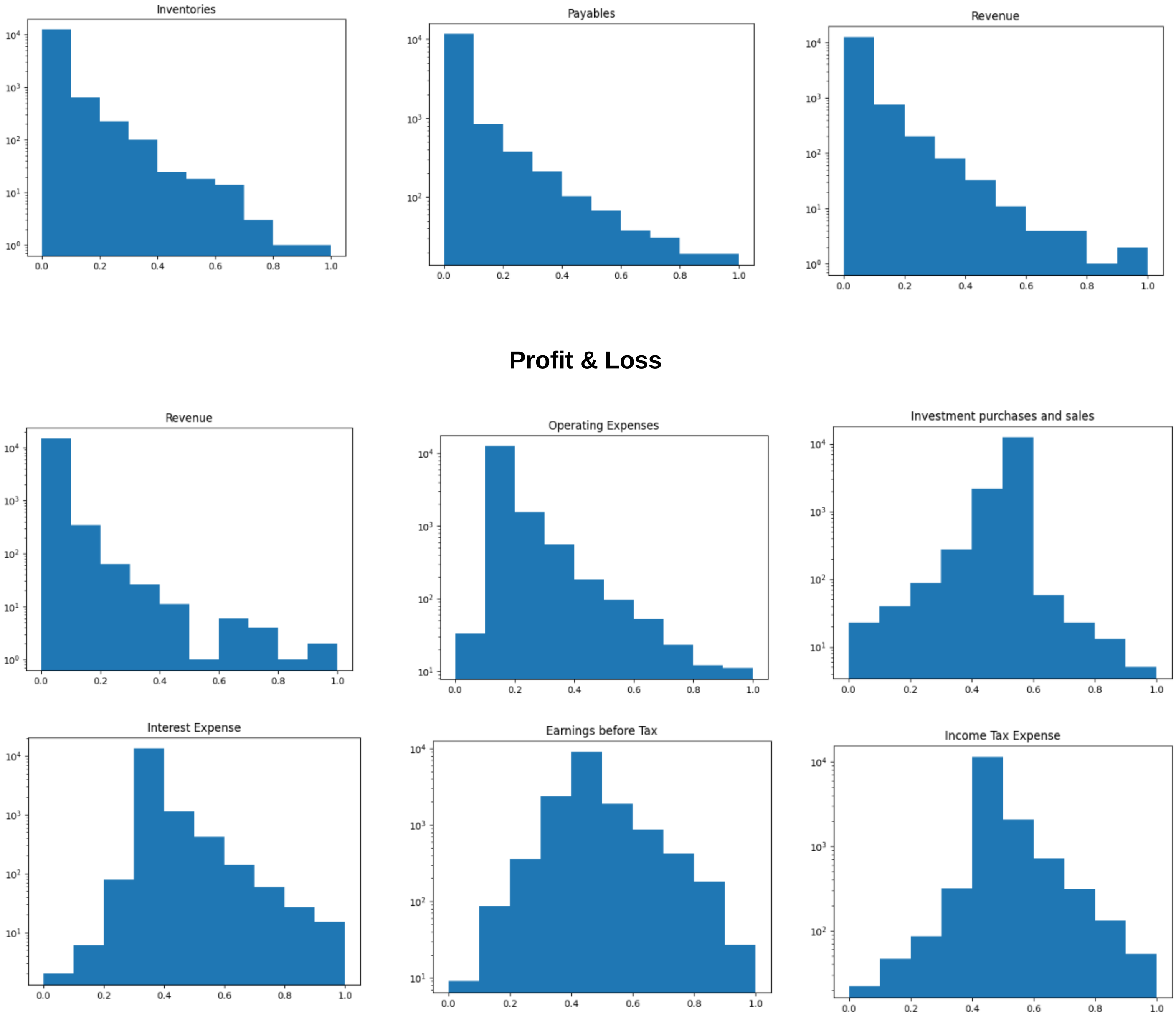
3.2.1. Distribution of Variables

The distribution of dependent and independent variables in the dataset is investigated by this study. The distribution of the 19 variables for the balance sheet and 16 for profit & loss in our dataset is plotted in histograms in Figure 1, which displays the findings. We can determine whether each variable is normally distributed through looking at the histograms in this display, which give an overview of each variable's distribution. For models like multiple linear regression, normality in all variables is not essential [26], but residuals must be normally distributed for all predictive models. A useful insight into the dataset that this analysis can offer is the relationships between the variables.

In this case, many distributions of variables are skewed to the right after removing outliers and normalizing the data, it means that most of the data points are concentrated towards the lower end of the range, with a few extreme values towards the higher end. This can affect the performance of some machine learning algorithms, as they may assume a normal distribution of data. In such cases, it may be necessary to apply additional transformations to the data, such as log transformation or power transformation, to make the distribution more symmetrical and improve the accuracy of the model. A sample of the distributions is shown below [Figure 3.2.1].

Balance Sheet



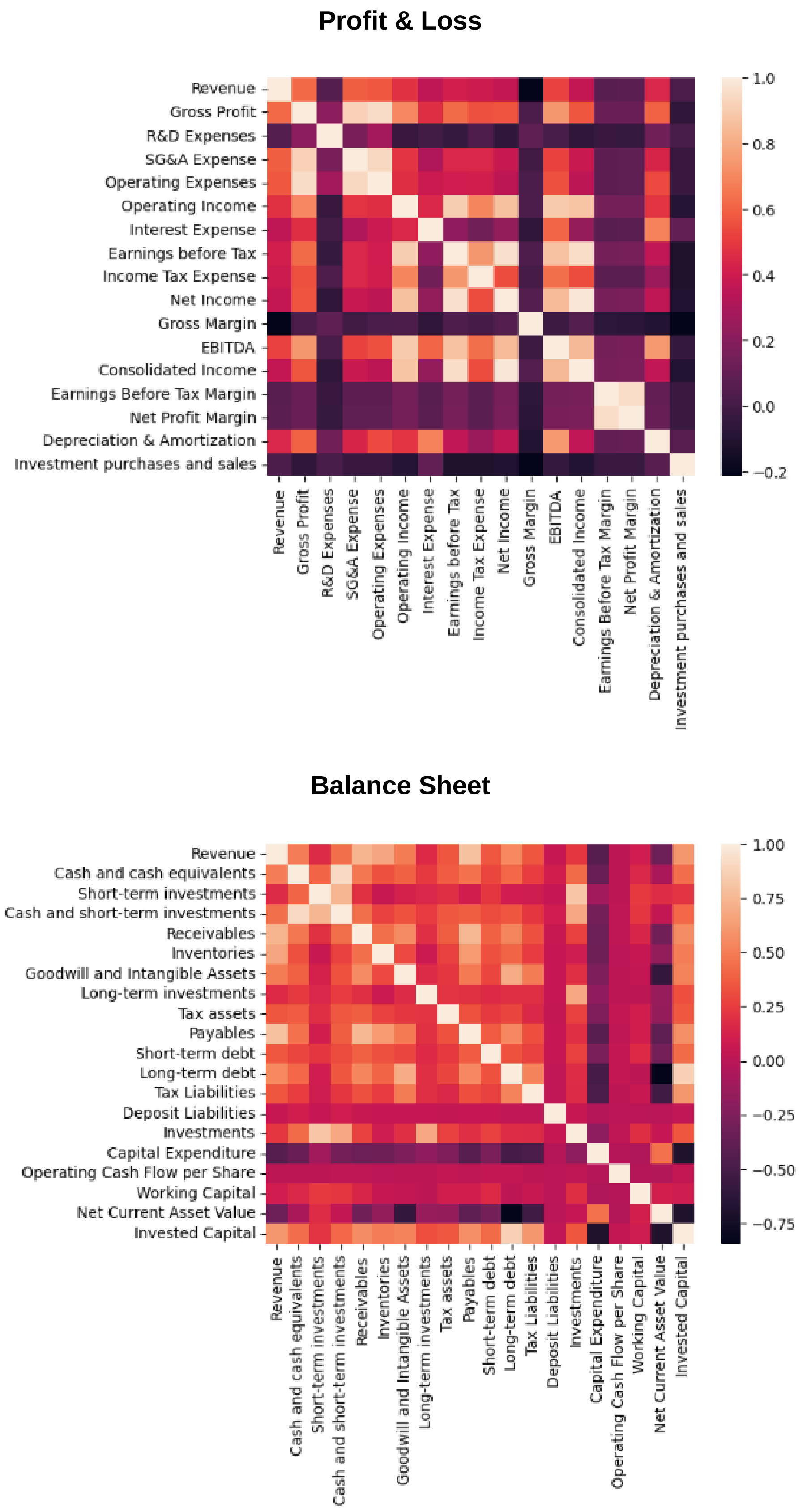


Profit & Loss

Figure 3.2.1: Distributions

3.2.2. Correlation of Variables

Checking the correlation between these variables is another method that can help uncover information about the variables in a collection. When two or more independent variables have a high correlation, this can result in multicollinearity, which is the existence of redundant predictors that have a high correlation [27]. In a statistical model, multicollinearity causes a number of problems, such as overfitting, unstable coefficient estimates, and trouble identifying important variables [28].



Profit & Loss

Balance Sheet

Figure 3.2.2: Correlations

A positive linear correlation value between two variables indicates that they move similarly, and a negative linear correlation value between two variables indicates that they move in opposition to one another, as is made abundantly obvious here. The correlation value is close to 0, which indicates that the patterns are not connected.

As we can see from the [Figure 3.2.2] above, there are several financial indicators that exhibit absolutely no linear correlation. Those are the heavily nan-dominant financial measures. This indicates that once the nan values are filled, the chart will alter.

By plotting once more the number of missing values and 0-valued entries that appear in the remaining financial indicators, we can assess the effect of our decisions in terms of threshold levels. Clearly, things have gotten better.

3.3. Modeling

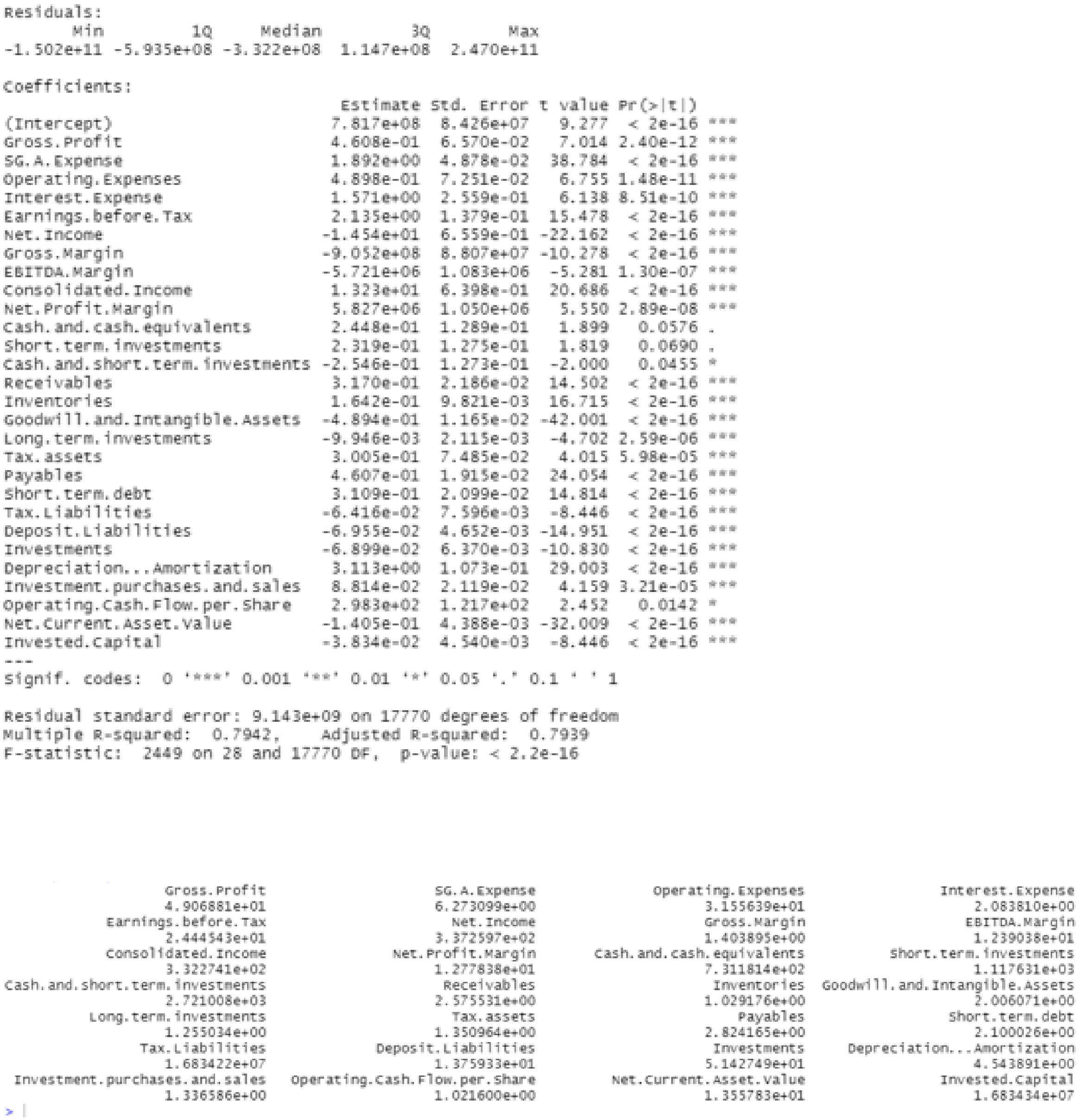
We will now concentrate on using the four machine learning methods that were discovered in the literature study. The two subcategories of these modeling strategies are variable ordering and variable selection. Our research includes LASSO and backward elimination for variable selection.

These techniques will be used on our dataset to produce a subset of independent variables that the model deems to be the most accurate at predicting the output of our dependent variable. Following that, we will rate the significance of each independent variable in our dataset using linear regression and random forest techniques.

3.3.1. Backfit

Backward elimination operates by evaluating all independent variables provided in a model and iteratively removing the least significant independent variables to the point that no insignificant independent variables are left [19]. This approach was selected because it can be easily understood and performs better than forward selection when there is collinearity between the independent variables [20].

To use this strategy, we will run a linear regression model first, and then backfit the linear regression model. Then, to ensure that the chosen independent variables are not strongly corelated, we will examine their VIF.



```
Residuals:
       Min         1Q     Median         3Q        Max 
-1.502e+11 -5.935e+08 -3.322e+08  1.147e+08  2.470e+11 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      7.817e+08  8.426e+07   9.277  < 2e-16 ***
Gross.Profit                     4.608e-01  6.570e-02   7.014 2.40e-12 ***
SG.A.Expense                     1.892e+00  4.878e-02  38.784  < 2e-16 ***
Operating.Expenses               4.898e-01  7.251e-02   6.755 1.48e-11 ***
Interest.Expense                 1.571e+00  2.559e-01   6.138 8.51e-10 ***
Earnings.before.Tax              2.135e+00  1.379e-01  15.478  < 2e-16 ***
Net.Income                      -1.454e+01  6.559e-01 -22.162  < 2e-16 ***
Gross.Margin                    -9.052e+08  8.807e+07 -10.278  < 2e-16 ***
EBITDA.Margin                   -5.721e+06  1.083e+06  -5.281 1.30e-07 ***
Consolidated.Income              1.323e+01  6.398e-01  20.686  < 2e-16 ***
Net.Profit.Margin                5.827e+06  1.050e+06   5.550 2.89e-08 ***
Cash.and.cash.equivalents        2.448e-01  1.289e-01   1.899   0.0576 .  
Short.term.investments           2.319e-01  1.275e-01   1.819   0.0690 .  
Cash.and.short.term.investments -2.546e-01  1.273e-01  -2.000   0.0455 *  
Receivables                      3.170e-01  2.186e-02  14.502  < 2e-16 ***
Inventories                      1.642e-01  9.821e-03  16.715  < 2e-16 ***
Goodwill.and.Intangible.Assets  -4.894e-01  1.165e-02 -42.001  < 2e-16 ***
Long.term.investments           -9.946e-03  2.115e-03  -4.702 2.59e-06 ***
Tax.assets                       3.005e-01  7.485e-02   4.015 5.98e-05 ***
Payables                         4.607e-01  1.915e-02  24.054  < 2e-16 ***
Short.term.debt                  3.109e-01  2.099e-02  14.814  < 2e-16 ***
Tax.Liabilities                 -6.416e-02  7.596e-03  -8.446  < 2e-16 ***
Deposit.Liabilities             -6.955e-02  4.652e-03 -14.951  < 2e-16 ***
Investments                     -6.899e-02  6.370e-03 -10.830  < 2e-16 ***
Depreciation...Amortization      3.113e+00  1.073e-01  29.003  < 2e-16 ***
Investment.purchases.and.sales   8.814e-02  2.119e-02   4.159 3.21e-05 ***
Operating.Cash.Flow.per.Share    2.983e+02  1.217e+02   2.452   0.0142 *  
Net.Current.Asset.Value         -1.405e-01  4.388e-03 -32.009  < 2e-16 ***
Invested.Capital                -3.834e-02  4.540e-03  -8.446  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.143e+09 on 17770 degrees of freedom
Multiple R-squared:  0.7942,	Adjusted R-squared:  0.7939 
F-statistic:  2449 on 28 and 17770 DF,  p-value: < 2.2e-16
```

```
                   Gross.Profit                    SG.A.Expense              Operating.Expenses                Interest.Expense 
                   4.906881e+01                    6.273099e+00                    3.155639e+01                    2.083810e+00 
            Earnings.before.Tax                      Net.Income                    Gross.Margin                   EBITDA.Margin 
                   2.444543e+01                    3.372597e+02                    1.403895e+00                    1.239038e+01 
            Consolidated.Income               Net.Profit.Margin       Cash.and.cash.equivalents          Short.term.investments 
                   3.322741e+02                    1.277838e+01                    7.311814e+02                    1.117631e+03 
Cash.and.short.term.investments                     Receivables                     Inventories  Goodwill.and.Intangible.Assets 
                   2.721008e+03                    2.575531e+00                    1.029176e+00                    2.006071e+00 
          Long.term.investments                      Tax.assets                        Payables                 Short.term.debt 
                   1.255034e+00                    1.350964e+00                    2.824165e+00                    2.100026e+00 
                Tax.Liabilities             Deposit.Liabilities                     Investments     Depreciation...Amortization 
                   1.683422e+07                    1.375933e+01                    5.142749e+01                    4.543891e+00 
 Investment.purchases.and.sales   Operating.Cash.Flow.per.Share         Net.Current.Asset.Value                Invested.Capital 
                   1.336586e+00                    1.021600e+00                    1.355783e+01                    1.683434e+07 
> |
```

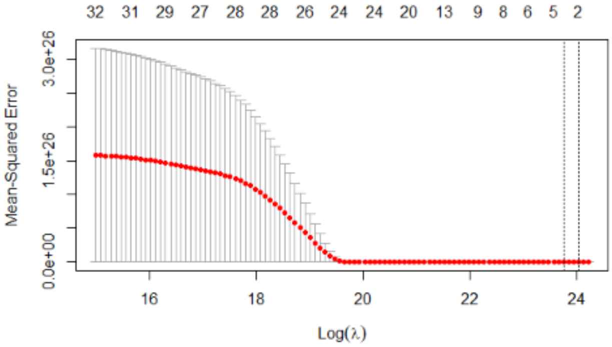
Figure 3.3.1: Backfit Summary and VIF

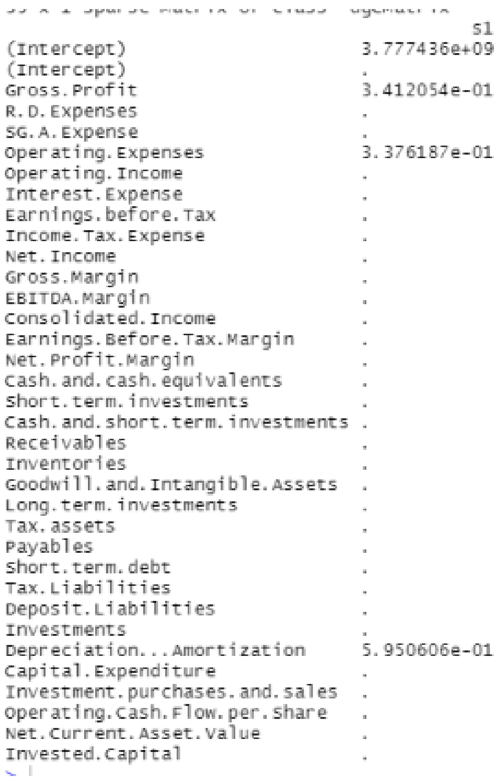
As seen in [Figure 3.3.1] these early findings make it clear that the backward elimination method chose a set of independent variables with a high degree of predictive power for the outcome of our dependent variable. The residuals are unbiased, all independent factors are statistically significant (p-value, which was set at.05), and the adjusted R-squared is 0.7939. The variables that have a VIF score above 5 indicate a strong correlation and are probably depicted in the regression coefficients.

3.3.2. LASSO

LASSO was also chosen as a feature selection technique for this study. This approach was chosen because it employs regularization, a technique that penalizes and reduces the coefficient of an independent variable to zero [21]. The production efficiency metric in this instance is the dependent feature, and this technique excludes variables it determines to be unimportant in forecasting the outcome of the dependent variable. Although LASSO appears comparable to other approaches, it was eventually selected because it makes use of both regularization and feature elimination.

We will apply the 'lambda' lambda value to this model. A LASSO model's regularization measure is called lambda, and the value "lambda" will produce the most regularized model, with an error that is no more than one standard error from the minimum [29].





```
35 x 1 sparse Matrix of class "dgCMatrix"
                                          s1
(Intercept)                     3.777436e+09
(Intercept)                     .
Gross.Profit                    3.412054e-01
R.D.Expenses                    .
SG.A.Expense                    .
Operating.Expenses              3.376187e-01
Operating.Income                .
Interest.Expense                .
Earnings.before.Tax             .
Income.Tax.Expense              .
Net.Income                      .
Gross.Margin                    .
EBITDA.Margin                   .
Consolidated.Income             .
Earnings.Before.Tax.Margin      .
Net.Profit.Margin               .
Cash.and.cash.equivalents       .
Short.term.investments          .
Cash.and.short.term.investments .
Receivables                     .
Inventories                     .
Goodwill.and.Intangible.Assets  .
Long.term.investments           .
Tax.assets                      .
Payables                        .
Short.term.debt                 .
Tax.Liabilities                 .
Deposit.Liabilities             .
Investments                     .
Depreciation...Amortization     5.950606e-01
Capital.Expenditure             .
Investment.purchases.and.sales  .
Operating.Cash.Flow.per.Share   .
Net.Current.Asset.Value         .
Invested.Capital                .
```

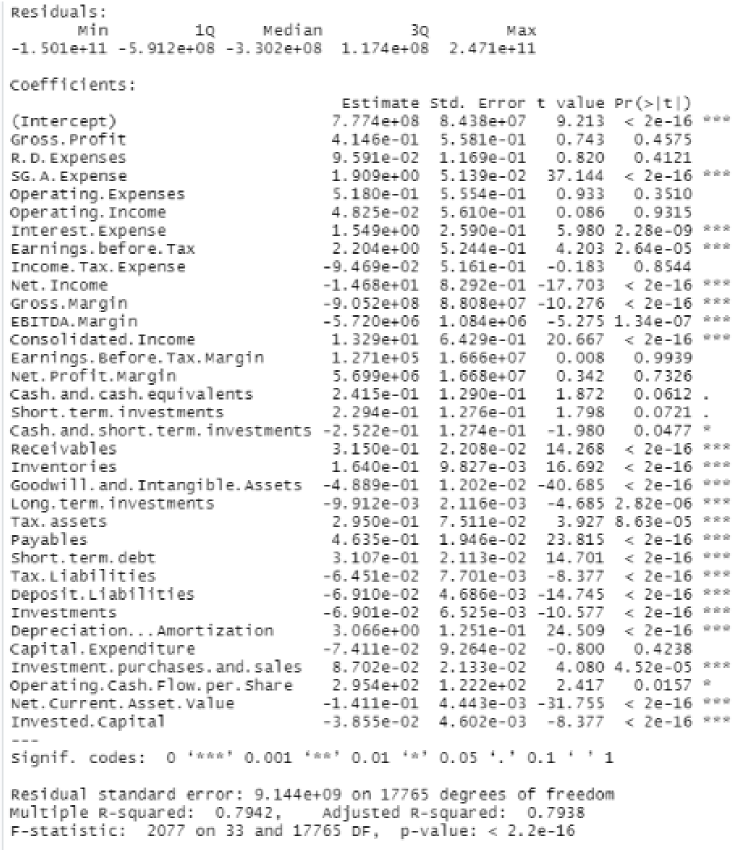
Figure 3.3.2: Lasso Summary

Despite the coefficients are different to the ones provided by the backward elimination model, it is heartening to note that the variables shared in the two models have the same effects on the dependent variable, indicating the two models can distinguish between positive and negative coefficients [Figure 3.3.2]. The outcomes of this model are thus likely to alter as a means of combining or removing these correlated metrics is discovered.

3.3.3. MLR

Professionals frequently use MLR because it generates a model output that is simple to understand and gives pvalues and coefficients to each independent variable [18].

We will use an MLR model in this modeling assignment to regress each of the independent variables against the dependent variable. After the model has been run, we will determine the ranking of variable significance according to the t-statistic's absolute value [Figure 3.3.2].



```
Residuals:
       Min         1Q     Median         3Q        Max 
-1.501e+11 -5.912e+08 -3.302e+08  1.174e+08  2.471e+11 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      7.774e+08  8.438e+07   9.213  < 2e-16 ***
Gross.Profit                     4.146e-01  5.581e-01   0.743   0.4575    
R.D.Expenses                     9.591e-02  1.169e-01   0.820   0.4121    
SG.A.Expense                     1.909e+00  5.139e-02  37.144  < 2e-16 ***
Operating.Expenses               5.180e-01  5.554e-01   0.933   0.3510    
Operating.Income                 4.825e-02  5.610e-01   0.086   0.9315    
Interest.Expense                 1.549e+00  2.590e-01   5.980 2.28e-09 ***
Earnings.before.Tax              2.204e+00  5.244e-01   4.203 2.64e-05 ***
Income.Tax.Expense              -9.469e-02  5.161e-01  -0.183   0.8544    
Net.Income                      -1.468e+01  8.292e-01 -17.703  < 2e-16 ***
Gross.Margin                    -9.052e+08  8.808e+07 -10.276  < 2e-16 ***
EBITDA.Margin                   -5.720e+06  1.084e+06  -5.275 1.34e-07 ***
Consolidated.Income              1.329e+01  6.429e-01  20.667  < 2e-16 ***
Earnings.Before.Tax.Margin       1.271e+05  1.666e+07   0.008   0.9939    
Net.Profit.Margin                5.699e+06  1.668e+07   0.342   0.7326    
Cash.and.cash.equivalents        2.415e-01  1.290e-01   1.872   0.0612 .  
Short.term.investments           2.294e-01  1.276e-01   1.798   0.0721 .  
Cash.and.short.term.investments -2.522e-01  1.274e-01  -1.980   0.0477 *  
Receivables                      3.150e-01  2.208e-02  14.268  < 2e-16 ***
Inventories                      1.640e-01  9.827e-03  16.692  < 2e-16 ***
Goodwill.and.Intangible.Assets  -4.889e-01  1.202e-02 -40.685  < 2e-16 ***
Long.term.investments           -9.912e-03  2.116e-03  -4.685 2.82e-06 ***
Tax.assets                       2.950e-01  7.511e-02   3.927 8.63e-05 ***
Payables                         4.635e-01  1.946e-02  23.815  < 2e-16 ***
Short.term.debt                  3.107e-01  2.113e-02  14.701  < 2e-16 ***
Tax.Liabilities                 -6.451e-02  7.701e-03  -8.377  < 2e-16 ***
Deposit.Liabilities             -6.910e-02  4.686e-03 -14.745  < 2e-16 ***
Investments                     -6.901e-02  6.525e-03 -10.577  < 2e-16 ***
Depreciation...Amortization      3.066e+00  1.251e-01  24.509  < 2e-16 ***
Capital.Expenditure             -7.411e-02  9.264e-02  -0.800   0.4238    
Investment.purchases.and.sales   8.702e-02  2.133e-02   4.080 4.52e-05 ***
Operating.Cash.Flow.per.Share    2.954e+02  1.222e+02   2.417   0.0157 *  
Net.Current.Asset.Value         -1.411e-01  4.443e-03 -31.755  < 2e-16 ***
Invested.Capital                -3.855e-02  4.602e-03  -8.377  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.144e+09 on 17765 degrees of freedom
Multiple R-squared:  0.7942,	Adjusted R-squared:  0.7938 
F-statistic:  2077 on 33 and 17765 DF,  p-value: < 2.2e-16
```

Figure 3.3.2: MRL Summary

3.3.4. RF

Random Forest is a well-known machine learning algorithm suitable for ranking variables by their significance. It can assess datasets of various sizes and model nonlinearities [18].

Gini importance and MSE reduction are two frequently used methods of ranking variables in Random Forest, which uses a tree-based model to build a regression function on a multivariate step function. Given that the Gini importance method is known to introduce biases, we will concentrate on the variable importance indicators generated by the MSE reduction methodology.

In contrast to MLR, parameters in Random Forest can be adjusted to best suit the model. The number of variables accessible for splitting at each tree node is known as mtry [30].

The model was able to reduce the OOB error by altering the mtry from the default option of the square root of the number of columns. We tried to improve the outcomes through trial and error because there was no easily oferred package to optimize the ntree parameter, which specifies how many trees the model should take into account. Ultimately, the default value of 500 was chosen.

3.4. Sensitivity Analysis

Sensitivity analysis is a technique to gauge the outcome variable's level of uncertainty by altering the input variables [31]. This analysis offers the chance to assess and comprehend how these factors relate to one another as a linear regression model. The creation of various scenarios that analyze the underlying situation when the input coefficient changes can also result from sensitivity analysis. Sensitivity analysis is frequently employed in a variety of sectors and academic fields.

Sensitivity analysis can be used, for example, to track how the predicting variable changes as the independent factors do. The model predicted that the output would become economically unviable over a specific number of the independent variable. The management teams can use this information to make excellent choices. The finished model can be used with similar strategies [32].

The effect of the chosen KPIs on our predicting output variable will be assessed first. The linear regression values can be interpreted to accomplish this. Second, each company can simulate various scenarios based on potential events using an optimization and simulation model to see how these scenarios affect the independent variables and, consequently, how the predicting variables and the efficiency measure interact. By creating a mitigation plan in light of these insights, we can better prepare for and respond to the possible events. Additionally, we can run simulations with various variable inputs to determine whether particular combos of the variable coefficients are practical.

Overall, the results of the sensitivity analysis will demonstrate how the outcome variable varies when one or more input variables are altered. From this point, businesses can create scenarios and simulations to improve their future planning while keeping the desired dependent output KPI [33].

3.5. Methodology Outline

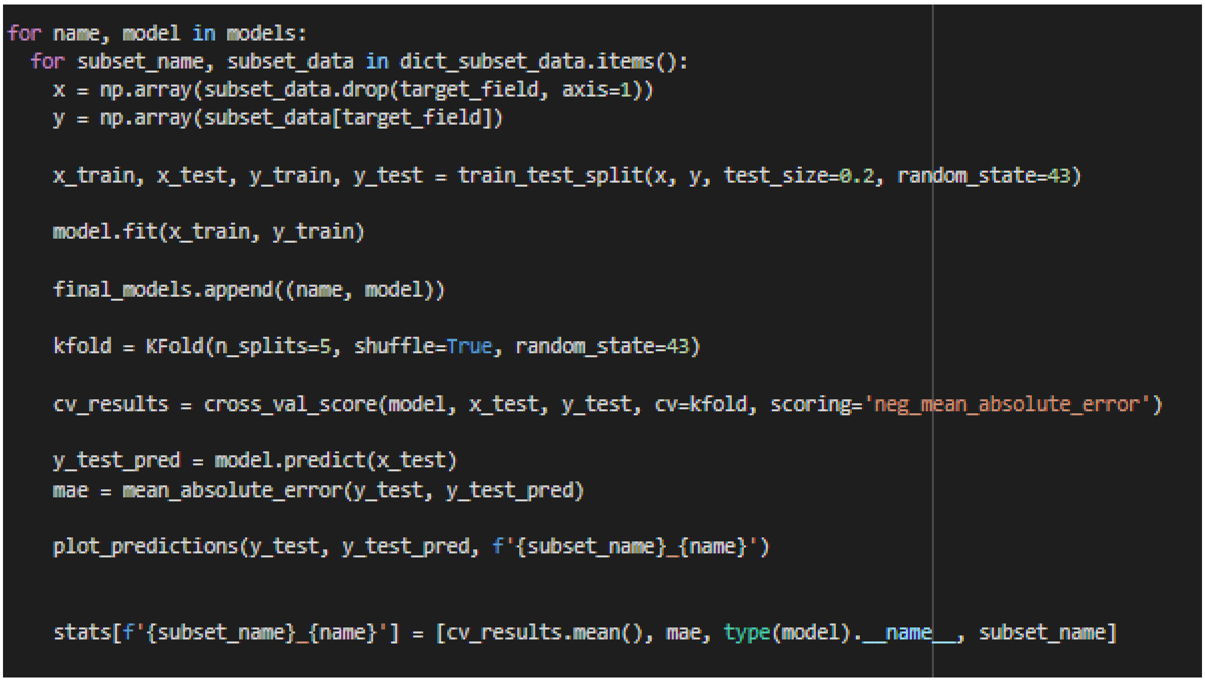
This section examined the dataset used, the way we processed the data, and the way used the various modeling approaches outlined in the theoretical background. Using this approach, we could identify the variables and characteristics that were most effective at forecasting the desired result for the company. We also discussed the sensitivity analysis that these models may be utilized to do in relation to the model's metrics. The best subset of KPIs for using multiple linear regression to forecast the dependent variable will be identified in the following chapter, along with a demonstration of how the business can benefit from the sensitivity analysis of the model.

4. RESULTS

This chapter starts out by going through the findings from all four models that were discussed in the implementation framework. The goal was to select independent factors for the final forecast model that are precise, succinct, yet straightforward to understand. For a variety of factors, the model valued these attributes. The model must, first and foremost, be able to forecast a company’s efficiency measure. KPIs that have no bearing on the dependent variable can be eliminated using a condensed collection of metrics. Finally, each company will have certainty regarding which metrics have the greatest influence on performance goals and the objectives that have to be established for each metric to achieve these goals by having an interpretable model.

4.1. Model Comparison

We will now compare the outputs of the four models in our study. In this stage, four independent linear regression models will be run for each of the features chosen by each method. The mean absolute error (MAE) of each model will then be evaluated in order to determine which set of KPIs is most accurate at forecasting the revenue efficiency measure. The MAPE formula divides the prediction for each time by the total number of individual absolute errors [34]. The average squared error's squared root is known as RMSE. Although we could calculate the accuracy of our predictions using either of these measures, we eventually chose to use MAE because it is straightforward and treats all residuals equally.



```
for name, model in models:
  for subset_name, subset_data in dict_subset_data.items():
    x = np.array(subset_data.drop(target_field, axis=1))
    y = np.array(subset_data[target_field])

    x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=43)

    model.fit(x_train, y_train)

    final_models.append((name, model))

    kfold = KFold(n_splits=5, shuffle=True, random_state=43)

    cv_results = cross_val_score(model, x_test, y_test, cv=kfold, scoring='neg_mean_absolute_error')

    y_test_pred = model.predict(x_test)
    mae = mean_absolute_error(y_test, y_test_pred)

    plot_predictions(y_test, y_test_pred, f'{subset_name}_{name}')


    stats[f'{subset_name}_{name}'] = [cv_results.mean(), mae, type(model).__name__, subset_name]
```

Figure 4.1.1: Training the models

As shown in [Figure 4.1] this code is performing machine learning modeling on a set of data subsets using different models. The first for loop iterates through a list of models and assigns each model to a variable called "model". The second for loop iterates through a dictionary of data subsets and assigns each subset's name to a variable called "subset_name" and the subset's data to a variable called "subset_data". The code then splits the subset data into training and testing sets using the train_test_split function from the scikit-learn library. The model is then fit to the training data using the fit method. The final_models list is then appended with a tuple containing the name of the model and the model object. The code then creates a KFold object with 5 splits and shuffles the data. The cross_val_score function from scikit-learn is used to calculate the mean absolute error (MAE) of the model's predictions on the testing data. The MAE is also calculated using the mean_absolute_error function from scikit-learn. The plot_predictions function is then called to create a plot of the actual vs predicted values of the testing data.

Finally, the stats dictionary is updated with the MAE, mean of cross-validation scores, model type, and subset name for each model and subset combination.

We can now build the four different predictive models after identifying our variable sets. We will calculate the MAE of each model after it has been constructed to determine which way of choosing KPIs had the least error in predicting the revenue KPI.

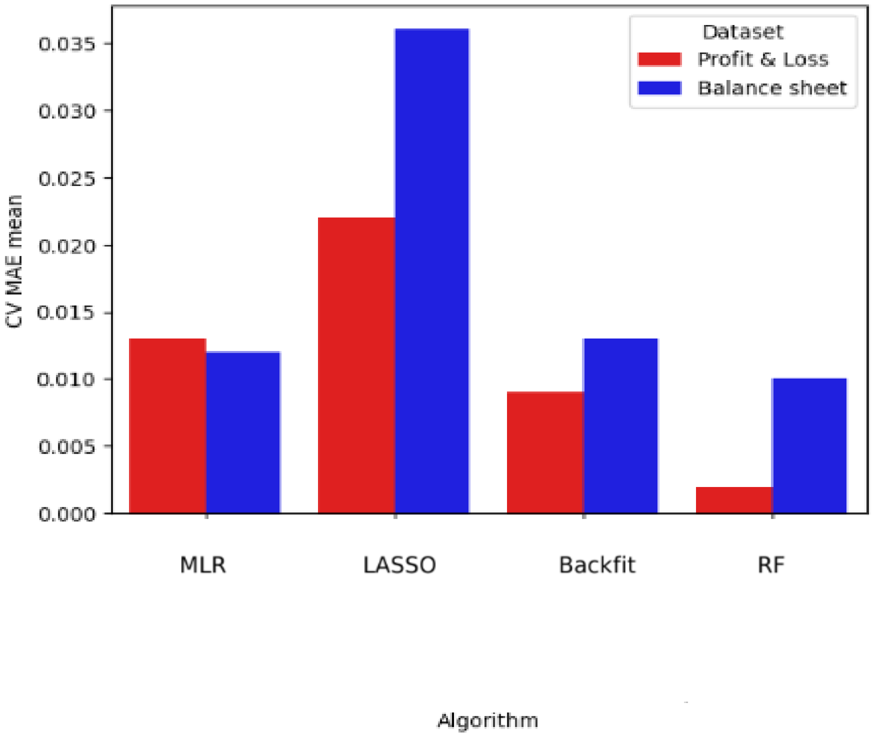


Figure 4.1.2: Models Results

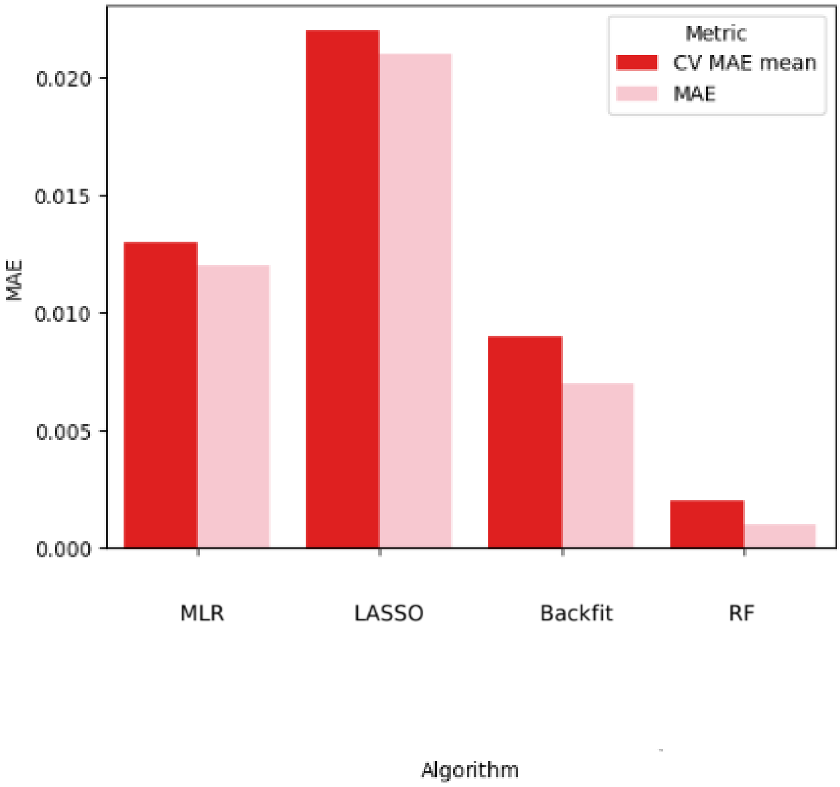
A few inferences can be made from these preliminary findings. First, from the selection models (backfit & LASSO) backfit produced lower MAE. From the ranking models (MLR, RF), Random Forest performed better (lower MAE). [Figure 4.1.2].

4.2. *Cross Validation*

We applied k-fold cross-validation for each of the models to verify and assess their performance. A common approach for testing machine learning models is this one. K-fold cross-validation works by dividing the provided dataset into k-folds, withholding one of the k-folds while training the model, and testing the trained model on the k-fold that was previously withheld. The same strategy is performed over randomly chosen subsets several times (referred to as repeats) employing recursive k-fold cross-validation, with the model's final error obtained as the mean of each repeat [35]. We chose this approach because it allows testing the models against numerous randomly generated test datasets, making it robust.

Prior to performing this validation technique, it was crucial to decide on the required number of k-folds and repeats. Throughout investigation, it was discovered that 5 or 10 k-folds are frequently chosen, as it has previously been demonstrated that these values produce outcomes without excessively high bias or variance [36]. We decided on 5 k-folds because it represents the smallest of the two options, and we have a relatively small amount of data records. We discovered that the cross-validation process frequently requires no more than ten repeats, so we set this parameter to ten [30]. The results are shown below. [Figure 4.2]

Profit & Loss



Balance Sheet

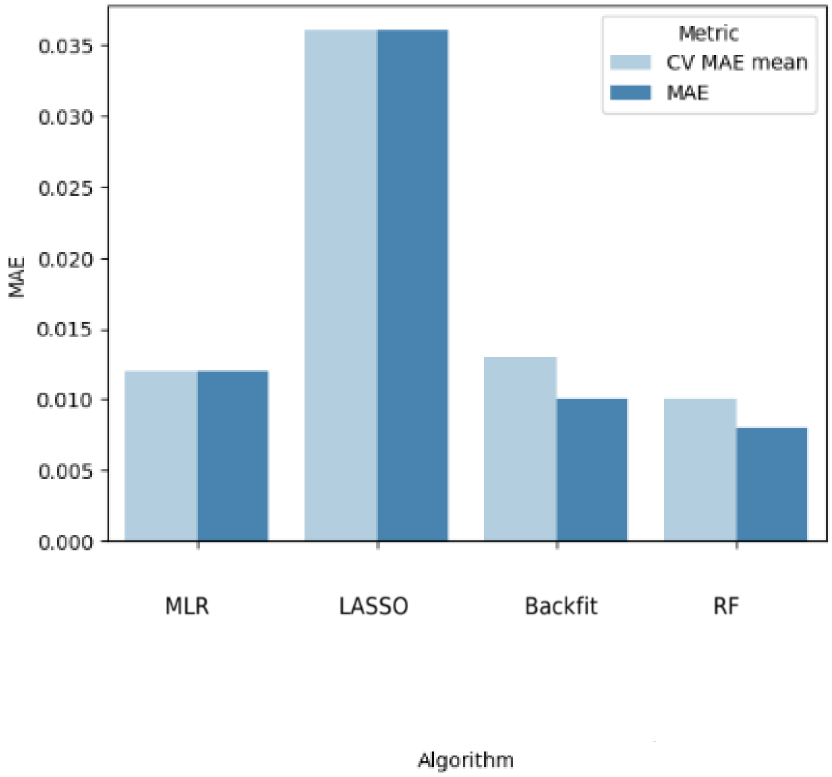


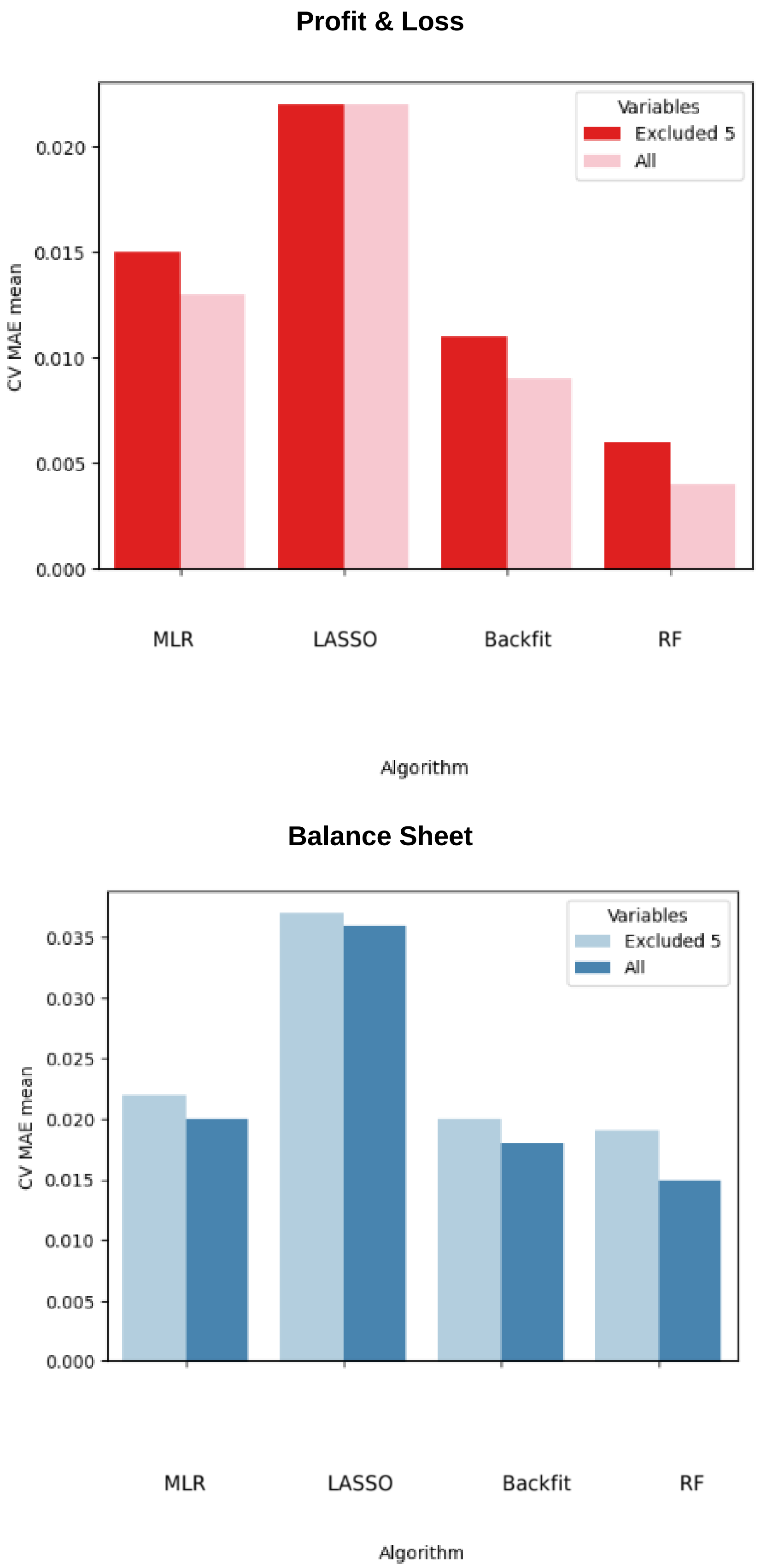
Figure 4.2: MAE Comparison

The figures show that the MAE for Profit & Loss has slightly increased for all four models. As new data is evaluated, models usually lose some of their predictive accuracy, so this was expected.

Even though multicollinearity has been cited as a cause for concern, this cross-validation analysis' results indicate the models chose KPIs that can accurately predict the results of the revenue.

4.3. Multicollinearity

We performed analysis again after realizing that there was multicollinearity between independent variables to check whether the results of the models would change if we removed five of the correlated variables from the dataset. We could see from the correlation diagram that they were highly correlated.


Profit & Loss

Balance Sheet

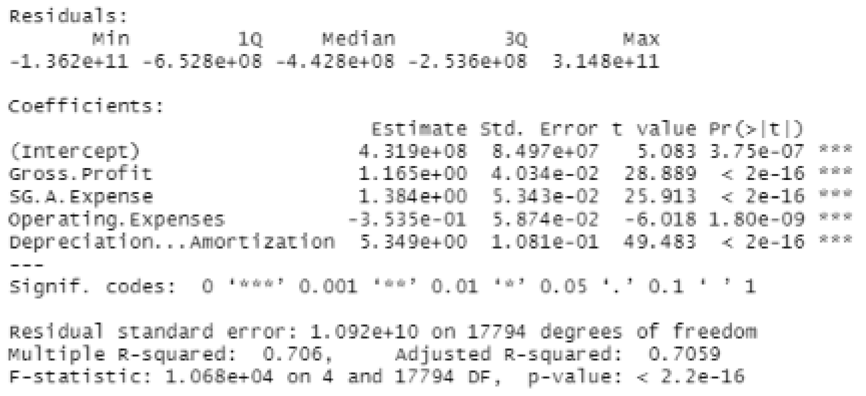
Figure 4.3: MAE Comparison with removed variables

[Figure 4.2] provides evidence in favor of the theory that multicollinearity had an impact on these models. The selection algorithms retain the highly correlated variables if every variable is available for the model, overfitting the models and lowering the MAE. When of them are present, multicollinearity also influences the coefficients and variable importance of the various variables in the MLR model, making the model less accurate. The models suggest resembling subsets of KPI measures with comparable forecasting abilities when either of the two variables with a high correlation is eliminated. When variables were removed from the dataset, RF algorithm again generated the lowest cross-validated MAE. According to these results all the models performed worse after the removal of the variables.

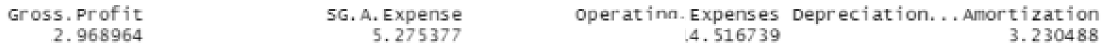
4.4. Selecting Key Performance Indicators

We will continue with the independent variables Gross Profit, SG&A Expense, Operating Expenses and Depreciation and Amortization and from Profit & Loss and Payables, Receivables and Inventories, for the final model and sensitivity analysis after taking into account all the data. With the help of the implementation analysis, the different statistical models were applied to the different versions of the variations of the dataset for forecasting the dependent variable, the most popular choices being Gross Profit, SG&A Expense, Operating Expenses and Depreciation and Amortization and from Profit & Loss and Payables, Receivables and Inventories

Profit & Loss

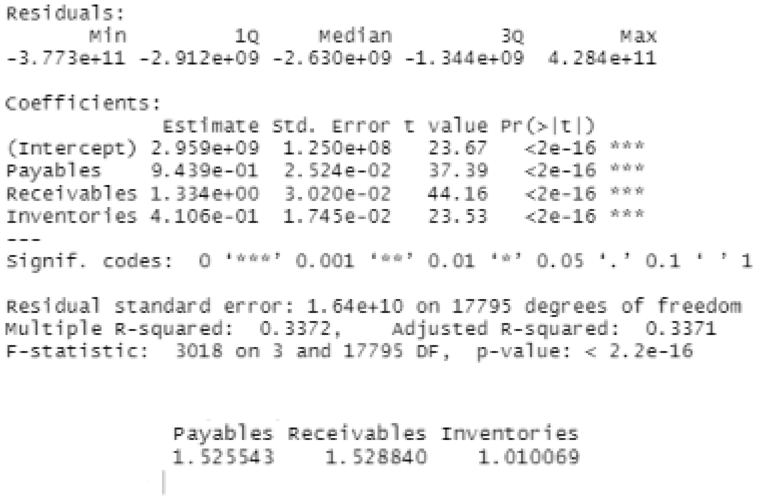


```
Residuals:
       Min         1Q     Median         3Q        Max
-1.362e+11 -6.528e+08 -4.428e+08 -2.536e+08  3.148e+11

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.319e+08  8.497e+07   5.083 3.75e-07 ***
Gross.Profit                1.165e+00  4.034e-02  28.889  < 2e-16 ***
SG.A.Expense                1.384e+00  5.343e-02  25.913  < 2e-16 ***
Operating.Expenses         -3.535e-01  5.874e-02  -6.018 1.80e-09 ***
Depreciation...Amortization  5.349e+00  1.081e-01  49.483  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.092e+10 on 17794 degrees of freedom
Multiple R-squared:  0.706,	Adjusted R-squared:  0.7059
F-statistic: 1.068e+04 on 4 and 17794 DF,  p-value: < 2.2e-16
```



```
Gross.Profit                SG.A.Expense          Operatinn.Expenses Depreciation...Amortization
    2.968964                    5.275377                    4.516739                    3.230488
```

Balance Sheet



```
Residuals:
       Min         1Q     Median         3Q        Max
-3.773e+11 -2.912e+09 -2.630e+09 -1.344e+09  4.284e+11

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.959e+09  1.250e+08   23.67   <2e-16 ***
Payables    9.439e-01  2.524e-02   37.39   <2e-16 ***
Receivables 1.334e+00  3.020e-02   44.16   <2e-16 ***
Inventories 4.106e-01  1.745e-02   23.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.64e+10 on 17795 degrees of freedom
Multiple R-squared:  0.3372,    Adjusted R-squared:  0.3371
F-statistic:  3018 on 3 and 17795 DF,  p-value: < 2.2e-16


             Payables Receivables Inventories
             1.525543    1.528840    1.010069
```

Figure 4.4: Final model summary

First, we determined that this group of KPIs does not contain any multicollinearity. The VIF score for all the variables is less than ten, indicating that multicollinearity is probably having an impact on them [37].

4.5. Insignificant Variables

Although this subset of KPIs and its subsequent multiple linear regression model have many positive quantitative and qualitative aspects, there are possible pitfalls. Although it is standard procedure to exclude variables that are not statistically significant, it is common to include this variable without negatively impacting the results of the final analysis when the evidence is sufficient.

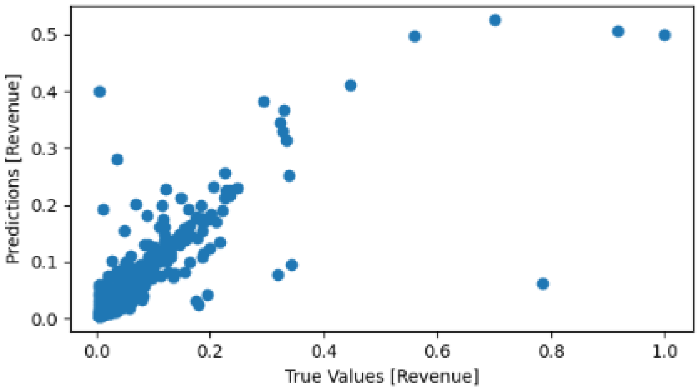
4.6. Output Prediction Results

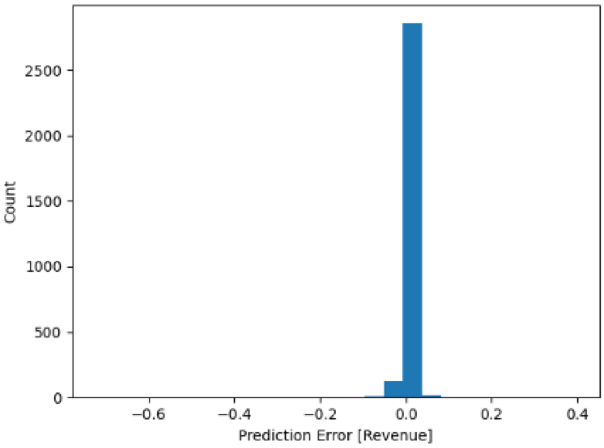
The prediction scatter plot and error distribution in machine learning show the relationship between the predicted values and the actual values of a model. The scatter plot displays the predicted values on the x-axis and the actual values on the y-axis, with each point representing a data point. The error distribution shows the distribution of errors between the predicted and actual values, with the mean and standard deviation of the errors displayed.

The scatter plot can help identify patterns or trends in the data, such as whether the model tends to overestimate or underestimate certain values. The error distribution can provide insights into the accuracy and precision of the model, as well as any potential biases or outliers in the data. Overall, these visualizations can help evaluate the performance of a machine learning model and identify areas for improvement.

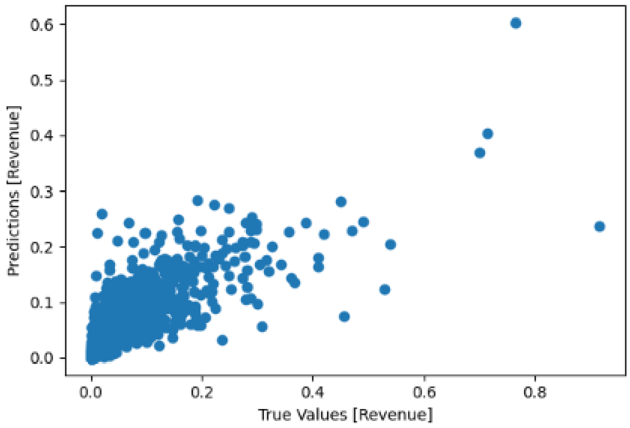
A good prediction should have a scatter plot with points that are closely clustered around the diagonal line, indicating a strong correlation between the predicted and actual values. The error distribution should be centered around zero, with a low variance and a normal distribution, indicating that the errors are small and evenly distributed. In general, a low mean squared error MAE value is also indicative of a good prediction. In our case the MAE for Profit & Loss is 0.004 and 0.05 for the Balance Sheet. By looking at the figures below [Figure 4.6] we can conclude that the prediction of the Revenue was successful.

Profit & Loss





Balance Sheet



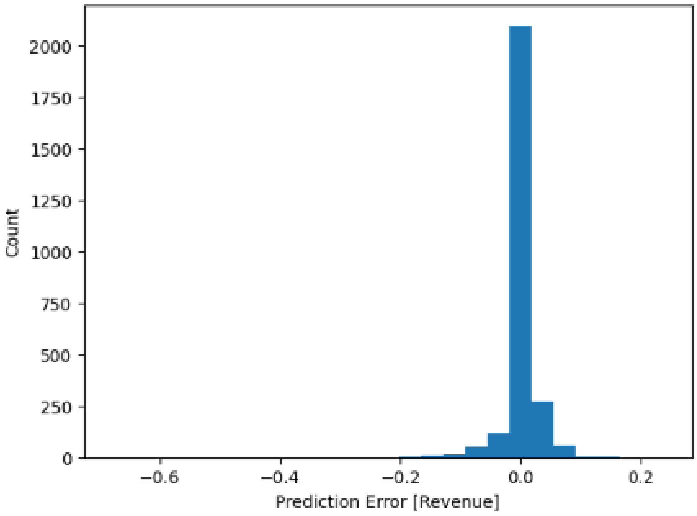


Figure 4.6: Revenue predictions and errors

4.7. Results Outline

To create a multiple linear regression model that could forecast revenue, we examined various combinations of the finance metrics offered within the dataset. Our findings indicated that Gross Profit, SG&A Expense, Operating Expenses and Depreciation and Amortization and from Profit & Loss and Payables, Receivables and Inventories were the most effective independent variables to forecast this output metric. The dependent variable was predicted using this subgroup of KPIs, which produced an adjusted R-squared of 0.7059 for Profit & Loss and 0.3371 for Balance Sheet. These findings indicate that using only these, more than 60% of the observed variation in the dependent variable can be explained for the Profit & Loss and more than 30% of the observed variation in the dependent variable can be explained for the Balance Sheet. It also enables the company to narrow its focus from the initial metrics included in the dataset to just these in order to manage its finances and meet its efficiency goals.

5. DISCUSSION

Determining how businesses could assess their existing KPIs throughout their finance activities to determine which parameters had the biggest effects on performance was the main goal of the thesis.

We also tried to figure out how these success parameters could give us information about goals that are in the future. We were able to satisfactorily respond to both questions after carefully examining the dataset.

We will go over the general importance of our analysis in this chapter, as well as how future experts inside any industry can use this approach to evaluate indicators and offer useful information to various businesses. We will discuss the benefits of the use of predictive analytics to evaluate finance parameters and the lessons learned by this approach, starting with managerial insights. Then, we go into great depth about how this framework can be used and the possible upgrades that may be achieved if this analysis was once again repeated.

5.1. Business Insights

It has been discovered from this study of the background that businesses evaluate metrics and KPIs as a typical practice for oversight of their performance. Although the techniques employed for the evaluations varied, it was discovered that most of the methods combined quantitative and qualitative analysis to identify the measures that were most valuable to every company. The suggested methodology of using predictive analytics offers a comparable degree of insight and is simpler to apply in today's business settings.

While they seek to make more informed choices across their organization, businesses worldwide are investing more and more in big data and business analytics. By 2025, the worldwide market for data and data analytics is expected to be worth $135 billion, according to Everest Group [38]. Companies are investing a great deal of money to oversee and gain knowledge from their data, which has led to a rising interest in data science within industry. Data scientists significantly improve decision-making, spot underlying trends and possibilities, and quantify business solutions for businesses by transforming data into business insights [39]. A great significance of using some of the instruments frequently used by data scientists was discovered, such as predictive analytics, to evaluate performance metrics, considering the increasing interest in data science and business analytics.

In addition, there are also useful justifications for using data science to evaluate performance. First, both academics and business frequently employ statistical modeling and predictive analytics. Predictive analytics models like multiple linear regression are ubiquitous in the curricula of many higher education programs [40]. Because such techniques are instructed frequently, it is possible that a certain individual inside a company has the knowledge to perform this analysis. It seems likely that several managers within a business are also acquainted with such techniques. A corporation or organization might be more inclined to require a third party to conduct the analysis. In addition, that they are free and accessible to to anybody with an electronic device and an access to the web, programming languages like R and Python are used to run these predictive models, making them a more accessible and affordable alternative to certain customized programs that may be required for complex linear programming models [16]. Due to these factors, the sole obstacle to using this method is obtaining the historical information required by the predictive models that monitor the performance metrics of an organization over time.

Applying our approach to evaluate measures using forecasting analytics is not just affordable and doable, yet it additionally provides insightful management information. Different performance metrics are frequently used today by different business divisions to oversee their staff and responsibilities.

Businesses can determine which metrics are most crucial to achieving their general objectives by using predictive analytics to evaluate operations metrics. A business can gain a lot from this analysis because it can pinpoint the metrics on which it should concentrate, get rid of KPIs that aren't improving performance, and identify areas where existing metrics could be improved. Additionally, it enables businesses to determine which of their KPIs are also KPIs for predictive analytics. Instead of traditional key performance indicators, that can frequently be centered on prior results, McKinney Rogers Ltd. claims that key predictive indicators assist managers in anticipating when they risk missing their long-term objectives [2]. When a business is aware of the metrics that are important in causing variance in its numerical goals, it can start to take proactive measures when one of its key predictive indicators starts to rise or fall.

Finally, our methods and findings have offered significant insights into management. Using machine learning and forecasting to evaluate finance metrics is simple to replicate, affordable, and successful at helping businesses get ready for the future. It is crucial to have the ability to quickly adapt and replicate analysis in a business setting that is constantly changing.

5.2. Further Analysis

We have developed a sound methodology for evaluating finance metrics using predictive analytics as a result of our study. Although this analysis was specifically designed for a specific, this implementation framework may be applied by experts and researchers working in a variety of fields. We will examine the applicability of similar analyses that may be used to analyze problems with similar problem formulations in this part. We will also point out the shortcomings of our research that could be improved by others in the future.

5.3. Interaction Effect & Sensitivity Analysis

In machine learning, an interaction effect refers to the situation where the effect of one variable on the outcome of a model depends on the value of another variable. In other words, the relationship between two variables is not additive, but rather multiplicative or nonlinear. For example, consider a model that predicts the likelihood of a customer buying a product based on their age and income. If there is an interaction effect between age and income, it means that the effect of age on the likelihood of buying the product depends on the customer's income level. In other words, the effect of age on the outcome is not constant across all income levels.

Interaction effects can be important to consider in machine learning because they can affect the accuracy and interpretability of the model. If an interaction effect is present but not accounted for in the model, the predictions may be inaccurate or misleading. Therefore, it is important to carefully analyze the data and consider potential interaction effects when building machine learning models.

Sensitivity analysis on the other hand is a technique used to determine how changes in one variable can affect the outcome of a model or decision. It is a powerful tool that helps decision-makers to understand the impact of different scenarios and make informed decisions. Sensitivity analysis involves varying one or more input variables while keeping all other variables constant and observing the resulting changes in the output. This analysis can be used to identify the most critical variables that affect the outcome of a model and to determine the range of values for these variables that will result in a desirable outcome.

In our case sensitivity analysis can be applied to determine the impact of different variables on the performance of the KPIs. For example, we may consider the impact of changes in interest rates, inflation rates, or exchange rates on the financial performance of the organization. By varying these variables and observing the resulting changes in the KPIs, we could identify the most critical variables that affect the financial performance of the organization and determine the range of values for these variables that will result in a desirable outcome. This analysis can help the organization to make informed decisions about its financial strategy and to develop a more robust KPIs framework.

5.4. Reproducing the Analysis

Defining company goals and gathering data are the first two steps in creating a predictive analytics model. Although they may appear simple, these two procedures serve as the foundation for implementing this approach. An individual must first comprehend the main target of the organization being researched and how to quantify it.

However, based on the company values, business life cycle, and general mission, other businesses' goals may differ from this one. For instance, if a new startup business wants to increase sales, one of its objectives would likely be revenue growth, that is usually assessed on a monthly, quarterly, or annual basis [41].

The following step will be to gather data that demonstrates the past patterns of this indicator in the business once a numerical target has been established. A company seeking to conduct this kind of evaluation ought to obtain all historical data for this indicator, to use revenue growth rate as an example. The dependent variable for the model will be this goal. The company must then determine all potential metrics or related datapoints that it thinks are having an impact on its numerical objective. These characteristics will function as independent factors in the model. These indicators or information must also be grouped at suitable amounts which correspond to the dependent variable, just like the objective value. To accurately evaluate the variance between these data points, for instance, each of the parameters should be grouped at the monthly basis using the objective of the monthly revenue growth rate. Since evaluating results over several years is frequent inside different sectors, grouping effectiveness over the years will probably be an extra frequent approach. [42].

The modeling process can start with a set of data that shows the past achievements of the company's numerical goal and associated KPIs and measures. According to our methodology, the company should evaluate the variables that are independent with respect to the dependent variable in two distinct manners: first, by using feature selection techniques to weed out the variables which are ineffective at forecasting the dependent variable, and secondly, by ranking the metrics' predictive importance to determine what indicators are the most crucial throughout the different groups. . The data analyst or investigator will be given various suggested subsets of measures to forecast the dependent variable by using these techniques. In this thesis, machine learning models were chosen to pick and rate metrics based on how well they matched the dataset supplied.

After gathering the data, a person must perform exploratory data analysis to comprehend the indicators contained in the set of data and afterwards choose the best models to forecast the results they are looking for. Any expert that aspires to duplicate this approach needs to assess various approaches and select the one that fits the specific set of data the best. Similar procedures should be followed to find an algorithm that is best suited to the particular dataset. It can be used as a starting point for anyone searching for a suitable variable ranking model. In conclusion, whichever algorithms are employed for the analysis, it is imperative that they be chosen in accordance with the dataset and issue statement provided.

The next stage is to interpret results and identify the best group of KPIs for sensitivity analysis after defining the objective and choosing the algorithms and models to rank and select KPIs. This guide of criteria for choosing this set of metrics should be that they accurately forecast the dependent variable, are succinct, and are simple to understand. The study ought to offer a subset of KPIs which only include pertinent indicators that are statistically established to influence the company’s goal, along with a numerical indication of the impact each KPI has on this goal, by adhering to the principles. These findings will serve as the foundation for sensitivity analysis, enabling leadership to comprehend the effects that changes in every measure might have on the company's overarching goals.

5.5. Potential Upgrades

While conducting this thesis, a few issues were found in the analysis that might be fixed if this approach was used in other contexts. First off, dataset didn't offer any detailed information about the dates connected to the data. Although was still possible to perform the analysis without these dates, having them might have given the models a boost from additional datasets. The setting required to perform time-series analysis and find any possible seasonality could have also been given by dates. To perform time series analysis, you need a dataset that includes a sequence of observations recorded at regular intervals over time. The data should be in chronological order, with each observation corresponding to a specific time period. The data should also be stationary, meaning that the statistical properties of the data do not change over time. For example, a company may collect data on sales over the past year to identify seasonal trends and adjust their inventory accordingly.

It is conceivable that a financial crisis, a natural disaster, or a worldwide pandemic could have an effect on a company's procedures. For example, a company may analyze the impact of changes in interest rates on their borrowing costs and adjust their financial strategy accordingly. If there are industry-specific trends that might be affecting finance, the corresponding stock prices of rival publicly traded companies could be used to determine this. These are just a few examples of the external data sources that might be used to uncover hidden correlations between finance measures and macroeconomic factors.

In summary, time-representative datapoints and datasets that provide context for external drivers have the ability to improve predictive modeling. Additional managerial insights can be gained by understanding how macroeconomic changes impact a company's operations. This knowledge can also help businesses develop more thorough sensitivity analyses to help them better prepare for the future.

5.6. Discussion Outline

We have come up with a thorough method for using predictive analytics to assess KPIs and performance measures. We discovered that using methods from this field to evaluate finance metrics is appealing for a variety of reasons, and that machine learning and data science are a set of skills that is rapidly expanding within the current business environment. This study was able to quantify the impact of KPIs on a company's overarching business goal and offer useful managerial insights into how these KPIs interact. This analysis also produced a framework for the proactive use of these metrics to help companies create business strategies by letting them know when they might be at risk of falling short of their intended goals. These findings show that our methodology can be applied to different sectors and that it can be enhanced with more data points.

6. CONCLUSION

Companies are working harder to develop their analytic skills in an effort to gain a competitive edge through better operations and consumer insights. Businesses all over the world are making significant investments in data analysis to gain knowledge from previous results, forecast the future ones, and simplify internal processes. Company managers nowadays depend on data to depend on analytics to determine, share information, and push their plan of action. This study concentrated on the use of KPIs in finance and the use of statistics and machine learning to help find the most appropriate indicators for a company to assess effectiveness right now and better plan for tomorrow.

A statistical modeling was postulated could aid a company in determining which indicators are most important for reaching a quantifiable goal and could assist in offering shed light on the company's potential future success. We chose and/or ranked the significance of KPIs using four techniques: Backfit, LASSO, MLR, and RF. The model with the greatest performance among these was Random Forest. However, given the problem statement, the need to narrow down the list of KPIs that have the greatest effect on the outcome variable, and the problem of the inventory KPIs' multicollinearity, we ultimately chose to use a Backward Elimination. We created a regression which produced an adjusted R-squared of 0.7059 for Profit & Loss and 0.3371 for Balance Sheet using the variables Gross Profit, SG&A Expense, Operating Expenses and Depreciation and Amortization and from Profit & Loss and Payables, Receivables and Inventories which produced an adjusted R-squared of 0.7059 for Profit & Loss and 0.3371 for Balance Sheet. By using this approach, a company may be in the position to comprehend how every metric interacts with the others. Additionally, by adjusting the variable inputs, they can optimize and simulate various situations to determine whether specific combinations of the metrics coefficients are a viable outcome.

The result of this methodology offers a comparable level of insight and is easier to reproduce in current business settings. Businesses are better able to comprehend the reasoning behind and create plans based on the model result as the multiple linear regression model is used more widely. The approach and model were in line with the premise, which held that statistical modeling could identify some of the crucial KPIs and direct tactical decisions.

A company can gain important insights by using predictive analytics to evaluate operations metrics because it can pinpoint the metrics it needs to pay attention to get rid of KPIs that aren't boosting performance, and find ways to make existing metrics better. Companies can also find the key predictive indicators that are important for influencing their performance using this strategy. Companies using this approach might be able to perform such modeling techniques again in the future to evaluate how their metrics affect operations change over time. The machine learning model is likely to become better and more useful for the company as more data and model versions are added. Anyone who replicates this analysis can also potentially strengthen their models and make changes by including information which gives perspective to external factors.

Although this study offered insightful analysis of the specific dataset, its real power rests outside the scope of this capstone. This approach can be used by many organizations and business functions because most big enterprises have access to historical data of their key indicators and a staff that is educated and competent in the specific techniques. Companies can determine the KPIs and macroeconomic factors which truly influence the success of their performance objectives by quickly analyzing historical performance trends. Additionally, they are able to recognize key predictors that can alert them to possible performance declines as specific circumstances arise. A company can also monitor whether these fresh efforts affect effectiveness over time by using critical learning indicators. Using these methods, management teams will be able to combine the KPIs which are presently using and might find new ones that the company can use. In conclusion, predictive analytics offer a quick, economical, and data-driven option to evaluating performance metrics.

REFERENCES

[1] Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., &amp; Sethupathy, G. (2016). The Age of Analytics: Competing in a Data-Driven World. Mckinsey Global Institute. 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

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

[3] Schrage, D.K. and M. (2019). Strategy For and With AI. [online] MIT Sloan Management Review. Available at: https://sloanreview.mit.edu/article/strategy-for-and-with-ai/.

[4] www.kaggle.com. (n.d.). 200+ Financial Indicators of US stocks (2014-2018). [online] Available at: https://www.kaggle.com/datasets/cnic92/200-financial-indicators-of-us-stocks-20142018.

[5] Twin, A. (2022). Understanding key performance indicators (KPI). [online] Investopedia. Available at: https://www.investopedia.com/terms/k/kpi.asp.

[6] 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.

[7] Shintaro, U., Shuangquan, S. and Satoshi, M. (2016). PSI-Cockpit: A Supply-Demand Planning System Focusing on Achieving Goals through Resolving KPI Conflicts. Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management, [online] pp.662–670. Available at: http://ieomsociety.org/ieom_2016/pdfs/182.pdf.

[8] ZHOU, H. and HE, Y.-L. (2018). Comparative Study of OKR and KPI. DEStech Transactions on Economics, Business and Management, [online] (eced). doi:https://doi.org/10.12783/dtem/eced2018/23986.

[9] Krueger, M., Luo, H., Ding, S.X., Dominic, S. and Yin, S. (2015). Data-Driven Approach of KPI Monitoring and Prediction with Application to Wastewater Treatment Process. IFAC-PapersOnLine, 48(21), pp.627–632. doi:https://doi.org/10.1016/j.ifacol.2015.09.596.

[10] 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.

[11] Franklin Liu, F.-H. . and Cheng Liu, Y. - (2008). Product line performance assessment on order fulfilment cycle time: a case of microelectronic communication company. International Journal of Production Research, 46(16), pp.4431–4443. doi:https://doi.org/10.1080/00207540600733543.

[12] Munier, N. and Hontoria, E. (2021). General Concepts. Management for Professionals, pp.1–4. doi:https://doi.org/10.1007/978-3-030-60392-2_1.

[13] Schmidt, K., Aumann, I., Hollander, I., Damm, K. and 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, [online] 15(1). doi:https://doi.org/10.1186/s12911-015-0234-7.

[14] 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.

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

[16] 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.

[17] Hayes, A. (2020). How Multiple Linear Regression Works. [online] Investopedia. Available at: https://www.investopedia.com/terms/m/mlr.asp.

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

[19] Heinze, G., Wallisch, C. and Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. Biometrical Journal, [online] 60(3), pp.431–449. doi:https://doi.org/10.1002/bimj.201700067.

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

[21] Fonti, V. (2017). Feature Selection using Lasso. VU Amsterdam. https://beta.vu.nl/nl/Images/werkstukfonti_tcm235-836234.pdf

[22] Nathan, N. (2019). Lasso, ridge and dropout regularization—Their effects on collinearity [online] Medium. Available at: https://towardsdatascience.com/different-forms-of-regularization-and-their-effects-6a714f156521

[23] Geisler Mesevage, T. (2021). What Is Data Preprocessing & What Are The Steps Involved? [online] MonkeyLearn Blog. Available at: https://monkeylearn.com/blog/data-preprocessing/.

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

[25] Roy, B. (2020). All about Feature Scaling. [online] Medium. Available at: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35.

[26] Bommae (2015). Should I always transform my variables to make them normal? | University of Virginia Library Research Data Services + Sciences. [online] Virginia.edu. Available at: https://data.library.virginia.edu/normality-assumption/.

[27] Hayes, A. (2022). Multicollinearity. [online] Investopedia. Available at: https://www.investopedia.com/terms/m/multicollinearity.asp.

[28] Wu, S. (2020). Multi-Collinearity in Regression. [online] Medium. Available at: https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea.

[29] Hastie, T., & Qian, J. (2023) glmnet.stanford.edu. (n.d.). An Introduction to `glmnet`. [online] Available at: https://glmnet.stanford.edu/articles/glmnet.html.

[30] Brownlee, J. (2016). Tune Machine Learning Algorithms in R (random forest case study). [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/.

[31] Chatterjee, S. and Hadi, A.S. eds., (1988). Sensitivity Analysis in Linear Regression. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:https://doi.org/10.1002/9780470316764.

[32] Abdallah, T., Diabat, A. and Simchi-Levi, D. (2011). Sustainable supply chain design: a closed-loop formulation and sensitivity analysis. Production Planning & Control, [online] 23(2-3), pp.120–133. doi:https://doi.org/10.1080/09537287.2011.591622.

[33] Kong,C. Charles, N. and Artman, S. (2021) ctl.mit.edu. (n.d.). Leveraging Predictive Analytics to Assess Operations Metrics. [online] Available at: https://ctl.mit.edu/pub/thesis/leveraging-predictive-analytics-assess-operations-metrics

[34] Vandeput, N. (2021). Forecast KPI: RMSE, MAE, MAPE & Bias. [online] Medium. Available at: https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d.

[35] Kassambara, A. (2018). www.sthda.com. (n.d.). Cross-Validation Essentials in R - Articles - STHDA. [online] Available at: http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validationessentials

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

[37] O'brien, R.M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), pp.673–690. doi:https://doi.org/10.1007/s11135-006-9018-6.

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

[39] Avantika Monnappa (2023). Why Data Science Matters And How It Powers Business in 2023? [online] Simplilearn.com. Available at: https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article.

[40] Murray, S. (2018) www.businessbecause.com. How Business Schools Are Teaching Big Data Analytics. [online] Available at: https://www.businessbecause.com/news/masters-in-business-analytics/5569/how-businessschools

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

[42] Likierman, A. (2014). The five traps of performance measurement. [online] Harvard Business Review. Available at: https://hbr.org/2009/10/the-five-traps-of-performance-measurement.