2022

Assessing predictive performance of supervised machine learning algorithms: an alternative model for diamond pricing.

Kigo, Samuel Njoroge
Strathmore Institute of Mathematical Sciences
Strathmore University

Assessing Predictive Performance of Supervised Machine Learning Algorithms: An Alternative Model for Diamond Pricing

Samuel Njoroge Kigo

Submitted in total fulfilment of the requirements for the degree of Master of Science in Statistical Sciences of Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

May 2022

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: **Samuel Njoroge Kigo**

Signature: 

Date: August 20, 2022

Approval

The thesis of Samuel Njoroge Kigo was reviewed and approved by the following:

Professor Bernard Omolo

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Evans Omondi

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

The world's hardest mineral is a diamond, which is 58 times harder than any other mineral, and its beauty as a jewel has long been appreciated. The diamond is popular due to its optical property as well as other causes such as its durability, custom, fashion, and strong marketing by diamond producers. Diamond demand, on the other hand, is not directly related to such inherent characteristics, but rather to their perceived value as rare and expensive objects. Forecasting diamond pricing is challenging due to non-linearity in important features such as carat, cut, clarity table, and depth. Given this, we conducted a comparative analysis and implementation of multiple supervised machine learning models in predicting diamond price in both classification and regression approaches. We evaluated eight different supervised algorithms in our work, including Multiple Linear Regression, Linear Discriminant Analysis, eXtreme Gradient Boosting, Random Forest, k-Nearest Neighbors, Support Vector Machines, Boosted Regression and Classification Trees, and Multi-Layer Perceptron, and showcased the best suitable model given selected evaluation metrics. The analysis in this work is based on data preprocessing, exploratory data analysis, training the aforementioned models, assessing their accuracy, and interpreting their results. Based on the performance metrics values and analysis, it was discovered that eXtreme Gradient Boosting was the most optimal algorithm in both classification and regression, with a R^2 score of 97.45% and an Accuracy value of 74.28%. As a result, the eXtreme Gradient Boosting method was recommended for forecasting the price of a diamond specimen.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ROC	Receiver Operating Characteristic	RMSE	Root Mean Squared Error
MAE	Mean Absolute Error	AUC	Area Under Curve
GIA	Gemological Institute of America	SMLAs	Supervised Machine Learning Algorithms
SML	Supervised Machine Learning	MLAs	Machine Learning Algorithms
ML	Machine Learning	BCART	Boosted Classification and Regression Tree
SVM	Support Vector Machine	MLP	Multi-Layer Perceptron
RF	Random Forest	kNN	K-nearest neighbors
XGBoost	eXtreme Gradient Boosting	MLR	Multiple Linear Regression
LDA	Linear Discriminant Analysis	ANN	Artificial Neural Network
ReLU	Rectified Linear Unit	ME	Misclassification Error

Acknowledgement

First and above all, I respectfully and gratefully recognize the almighty God for the gift of grace and capacity, as well as the gift of wisdom and knowledge, that He has bestowed upon me in order to conduct this research. Prof. Bernard Omolo and Dr. Evans Omondi, my supervisors, deserve special thanks for their unwavering support and advice during the research time. Finally, I want to express my gratitude to my parents, Mr. and Mrs. Kigo, brother Robert Kigo, nephew Kennedy Mwangi, and all of my fellow classmates (Kroneckers) for their essential moral and technical assistance throughout this thesis.

Dedication

Mr. Paul Gitau, my mentor and sponsor, is honored with this thesis. In the last decade, he has been my mentor, and over that time, I have felt a deeper feeling of intellectual worthiness in my life. His guidance has aided me in defining and even exceeding my personal boundaries. He has always pushed my academic abilities, knowing full well that I would always do better. This has been my compass for reaching my potential, and it has enabled me to achieve more than I could have dreamed, particularly in my academic career. May the Almighty bestow His blessings on him.

Chapter 1

Introduction

1.1 Background of the Study

A diamond is the hardest substance on the planet, and its beauty as a gemstone has long been recognized. By 2019, it was estimated that 142 million carats of diamonds have been mined around the world. Australia, Canada, the Democratic Republic of Congo, Botswana, South Africa, and Russia are all major producers. There are an estimated 1.2 billion carats in the world's reserves. The largest reserves are in Russia, which are believed to be worth 650 million carats (M.Garside, 2022).

In 2020, global diamond jewelry sales were 68 billion (in US dollars) according to M.Garside (2021b), with the United States accounting for 35 billion (in US dollars) of that amount (M.Garside, 2021a). In 2019, the United States had the biggest demand for polished diamonds, totaling 12.8 billion (in US dollars) according to (M.Garside, 2020). The United States, China, India, Japan, and the Persian Gulf region are the top five markets for diamonds, according to (Mamonov and Triantoro, 2018).

The 4Cs - Cut, Carat, Color, and Clarity - were introduced by the Gemological Institute of America (GIA) in the 1950s and are the most well-known attributes of diamonds. The 4Cs describe each diamond's distinct characteristics and have a significant impact on diamond prices. Three of the four Cs have a lengthy history: carat weight, color, and clarity were all utilized in the original diamond grading system over 2,000 years ago in India (Mamonov and Triantoro, 2018).

The dimensions of a diamond's cut determine how efficiently it reflects light. On a scale of fair to ideal, the cut perfection is classified.

One of the main components of the cut variable is the degree of perfection achieved by the cutting and polishing process. This is a complicated variable that includes, among other things, the stone's symmetry and adherence to local market-specific standards regarding the stone's proportions and the presence or absence of specific features such as an ID number engraved in the stone's girdle, the girdle's faceting, and so on (Cardoso and Chambel, 2005).

The cut of a diamond also has three other characteristics: brilliance, or the amount of light reflected from it; fire, or the dispersion of light into the colors of the spectrum; and scintillation, or the flashes of glitter that occur when a diamond is moved around (Mamonov and Triantoro, 2018). According to the International Gem Society, out of 4Cs, the cut is the most important attribute of a diamond (Clark, 2022). Chu (2001) asserts that optimal cut should neither be too deep nor too shallow for it will impede the trajectory of light and thereby the brilliance or "fire" of a diamond stone.

Blue Nile, one of the largest online diamond retailers, asserts that cut has the biggest effect on the sparkle, and even with perfect color and clarity, a poorly cut diamond will look dull (Nile, 2022). In addition to 4Cs, there are many other attributes of diamonds such as length, width, height, table, etc. To better understand how such complex features influence diamond prices, we propose application of supervised machine learning algorithms (SMLAs). SMLAs afford the advantage of capturing non-linearity relationships in a given dataset.

Buyers and investors in the diamond trade industry encounter a number of challenges in estimating the price of diamond stones. Because of the differences in the shapes, sizes, and clarity of the stones, this is a challenging task (Alsuraihi et al., 2020). Diamonds are a very unique consumer product. It is the hardest mineral i.e. 58 times harder than any other mineral, Mihir et al. (2021), thus used in various machines and other types of equipment for cutting and slicing.

Diamonds demand, on the other hand, is not directly tied to such inherent features, but rather to their perceived value as rare and expensive items (Mamonov and Triantoro, 2018). It is one of the gemstones on which more money is spent than any other combined gemstone.

The diamond gains popularity because it has an optical property (Mihir et al., 2021). Other factors include its durability, custom, fashion, and aggressive marketing by diamond producers (Sharma et al., 2021). Due to its non-linearity and fluctuating time series behavior, forecasting the prices of precious metals such as gold and diamonds is a difficult task (Mamonov and Triantoro, 2018).

Because of their unique features and great market demand, banks prefer to invest in precious metals. As a client, you're constantly unsure when it's the best moment to invest in, buy, or sell valuable items like gold and diamonds (Pandey et al., 2019). When it comes to generating the greatest profit out of an investment and the least expense out of a purchase made for the above-mentioned products, pricing is extremely important to buyers and investors.

Rapaport list, price list that quote Rapaport's opinion of high cash asking prices generally accepted as high enough to serve as the initial starting point for negotiations, has been used in classical diamond evaluation. The Rapaport list prices are almost always higher than actual dealer transaction prices, which tend to trade at discounts to the list. Final transaction prices are the result of negotiations between buyer and seller thus being difficult to predict based only on the Rapaport price list (Cardoso and Chambel, 2005).

There are numerous models and applications currently in the market which are used for predicting the future price of diamond stones, including machine learning algorithms Sharma et al. (2021), as well illustrated in section two of this paper.

Contemporary statistical analysis is characterized with the evolution of Machine Learning Algorithms. Ahmed et al. (2010) and Kampichler et al. (2010) observe that these algorithms have been empirically proven to be serious contenders to Classical Statistical Models in dealing with high dimensional data that are often non-linear and do not meet the assumptions of conventional statistical procedures. This aspect thus affirms the decision to employ them in this work for diamond price prediction and classification.

The aforementioned assertions are based on various predictive performance metrics including Precision, Accuracy, Kappa, Recall, F-Measure, RMSE, MAE, R Squared and computational aspects such as speed and build time, among others.

The knowledge of best performing model(s) is imperative in refocusing the modeler's time, effort and other resources to only potential candidate models in machine learning and pattern recognition.

In classical statistics, modeling complex non-linear relationships was the biggest drawback until 1980's where advancement in computing technology permitted non-linear modeling (James et al., 2013). The explosion of 'Big Data' has seen the release of over 95% of the current world data in just about the past five years. There exists a pressing need by businesses, governments and researchers to draw meaningful insights out of these overwhelming amounts of data in making smarter decisions.

In fact, businesses are now viewing data as a cocktail of new opportunities or a crude oil that requires state-of-the-art skills and expertise to refine, i.e. gaining insight into the engineering process behind data thus discovering the hidden patterns that will inform valuable versions of investment decisions. It is evident that data has become the new front of competition among businesses and other commercial underpinnings.

The new epoch of big data is redefining statistical learning applications on supervised and unsupervised modeling and prediction. Osisanwo et al. (2017) postulate that this tendency can be traced back to the advancement of Smart and Nano technologies, which has sparked interest in uncovering hidden patterns in data, both structured and unstructured, to derive value. Further, increase in the freely available and user-friendly statistical softwares such as *R* and *Python* has provided an upthrust to machine learning innovations in modeling.

In this paper, we explore the use of supervised machine learning algorithms to investigate the relationship between diamond physical qualities and diamond prices in order to establish the extent to which the latter are determined by the former.

1.2 Problem Statement

Cut is the most important variable influencing diamond prices, according to the various literature materials reviewed. The cut itself contains three key aspects: brilliance, dispersion, and scintillation, all of which attract the attention of the diamond market's major players.

Thus, in addition to predicting diamond prices, we capture the aspect of diamond classification based on cut in this study, which has received little to no attention in previous research. As a result, the predictive power of SMLAs in predicting diamond prices and diamond classification based on cut will be evaluated in this paper. An algorithm's overall performance will be judged depending on how well it performs in both regression and classification scenarios.

To address the said gap, we employ the following Supervised Machine Learning Algorithms: boosted classification and regression trees (BCART), support vector machines (SVM), Multi-Layer Perceptron (MLP), random forest (RF), K-nearest neighbors (KNN) and eXtreme Gradient Boosting (XGBoost). Multiple Linear Regression (MLR) and Linear Discriminant Analysis (LDA) from classical statistics will be used as baseline models.

Though not the main subject, the study will delve into ensemble methods, defined as a collection of models whose predictions are combined by weighted averaging (continuous response variable) or plurality voting scheme (categorical response variable) (Moisen, 2008). On this objective, a boosted or bagged version of an algorithm is expected to perform better than the baseline counterpart.

Vafeiadis et al. (2015) postulate that boosting ameliorates the performance of a classifier based on the respective F-measure score. Despite the fact that ensemble approaches have outstanding empirical performance, Bucilua et al. (2006), most model comparison studies have not applied them.

1.3 Objectives of the Study

1.3.1 General Objective

The general objective was to assess predictive performance of supervised machine learning algorithms on diamond pricing.

1.3.2 Specific Objectives

- To assess the predictive performance of Supervised Machine Learning Models in diamond price prediction.
- To assess the predictive performance of Supervised Machine Learning Models in diamond classification.
- To compare the performance of boosting and bagging (bootstrapped aggregation) in ensemble methods.

1.4 Research Questions

This study ought to answer the following questions:

- What is the predictive performance of Supervised Machine Learning Models in diamond price prediction?
- What is the predictive performance of Supervised Machine Learning Models in diamond classification?
- How does the performance of boosting and bagging (bootstrapped aggregation) compare in ensemble methods?

1.5 Significance of the Study

The findings will have larger ramifications for how online commerce affects the pricing of diamonds and other luxury products. Furthermore, the results will have an impact on future strategies for the diamond industry's major players, such as addressing the pricing pressure imposed by e-commerce. We propose an alternate approach as demonstrated by the 'overall modeling process' in chapter three, based on the diamond cut, a feature that can be easily tweaked by dealers to meet market demand while also maximizing earnings and customer satisfaction.

1.6 Dissemination and Utilisation of the Study Results

The findings will be disseminated through publications and the Strathmore University Library Catalogue. The important actors in the diamond market, i.e. suppliers and purchasers for whom pricing is critical, will be the consumers of the research outputs. This is intended to create a near-ideal market in which both buyers and sellers have access to the same information. The optimal model will be based on an interactive space, such as R-Shiny, where diamond attributes are fed and the model generates the most accurate price estimate.

1.7 Limitations of the Study

This analysis is predicated on the premise that diamond unit prices will not fluctuate much over time, as this would make the model unstable. Diamond prices, on the other hand, appear to be rather steady throughout time, at least in the short to medium term.

Chapter 2

Literature Review

2.1 Introduction

This section provides a summary of the numerous sources that were studied for this work, including books, scholarly articles, and other materials. To create the groundwork for the research topic, a critical examination of the materials was carried out.

2.2 Supervised Machine Learning Algorithms

The urge to analyze data with most efficient Machine Learning Models has provided a springboard to a proliferation of model comparison studies in the past decade. However, the application of Machine Learning Algorithms (MLAs) demand a wide array of skills, most of which are not within the scope of many practitioners (Vafeiadis et al., 2015). This fact has motivated most scholars in statistics domain, including this research, to add more knowledge to the MLAs' bank through publications.

Supervised machine learning (SML) refers to the quest for algorithms that reason from externally supplied instances to develop general hypotheses, which then make predictions about future instances, based on certain intelligent systems (Osisanwo et al., 2017).

According to James et al. (2013), SML consists of building mathematical models for predicting the outcome of future observation. Predictive models can be classified into two main groups: *regression analysis* for predicting a continuous variable and *classification* for predicting the class or group of individuals.

2.3 Application of ML in Classification and Regression

Caruana et al. (2008) assert that most ML models' comparison studies have exclusively and extensively focused on classification problems. Vafeiadis et al. (2015) evaluate performance of five most widely employed classification algorithms on customer churn predictions. Song et al. (2004) compare ML classifiers against classical statistical classification models. Recently, few studies have however been founded on regression problems for example Phaladisailoed and Numnonda (2018) which is meant to predict bitcoin prices and Salazar et al. (2015) that employs machine learning techniques to study dam behavior.

2.4 Application of ML in Diamond Pricing

Alsuraihi et al. (2020) seeks to develop the best algorithm for diamond dealers to employ in order to accurately estimate prices. Due to the wide diversity in diamond stone sizes and other important factors, the prediction procedure is much more challenging in the case of diamonds.

Several machine learning methods, including Liner regression, Random forest regression, polynomial regression, Gradient descent, and Neural network, were utilized in this article to aid in the prediction of diamond price. After training numerous models, verifying their accuracy, and analyzing the findings, it was discovered that the random forest regression is the best, with MAE and RMSE values of 112.93 and 241.97, respectively. Although RF performed exceptionally well in this study, a setting with a significant class imbalance which was the case for the dataset under analysis is not appropriate for this algorithm (Wu et al., 2014).

Importantly, the study should have taken into account classification, which is a significant factor in diamond pricing. The diamond cut, which comes in five different types (classes), has a significant impact on the price. As a result, the study should have chosen the optimal model based on classification and regression findings to capture this noble attribute. Finally, the study may have used a combination of ensemble models to improve the results.

Mamonov and Triantoro (2018) establishes that e-commerce has made it easier for buyers to compare diamond pricing (price dispersion) with diamond physical qualities across different sellers in order to make informed purchasing selections. The purpose of this study is to investigate the relationship between diamond physical features and diamond prices in order to identify the extent to which physical attributes influence diamond pricing.

The primary variables that determine diamond prices are discovered to be diamond weight, color, and clarity. The data mining findings also point to a significant level of subjectivity in diamond pricing, which could be due to diamond dealers' price obfuscation methods. The newly discovered information contributes to our knowledge of the relationship between consumer search costs and price volatility.

Because diamond price is a continuous interval target variable with a ratio scale, decision forest, boosted decision tree, and artificial neural network prediction data mining approaches have been used. When the complete dataset is analyzed, Decision Forest yields the lowest MAE, 5.8 percent. When the carat range in the diamonds dataset is narrowed to 0.2 - 2.5, ANN achieves an MAE of 8.2 percent, beating other techniques.

Other novel prediction data mining approaches, such as XGBoost, SVR, Knn, and others, which have been empirically demonstrated to produce good outcomes in model comparison studies, have been left out of this work for no obvious reason. Given the non-linearity of the variable relationship, SVR's polynomial option (svmPoly) may have produced superior results.

It's possible that a stacking ensemble with a one-time model run was used. R^2, RMSE, and other evaluation criteria were not employed. Despite the fact that the research suggests that the cut is the most important of the 4Cs, it is not included in the analysis, for example when trying to classify diamonds based on this attribute. The cut of a diamond determines its market value in this case.

Pandey et al. (2019) states that diamond and precious metal values fluctuate on a regular basis, making it difficult to forecast future value. This study uses ensemble approaches to anticipate future prices of precious metals such as gold and precious stones such as diamonds,

with the goal of obtaining the most accurate result possible. Also employed are feature selection approaches, and the outcomes are compared.

Over-fitting and under-fitting are common problems with supervised models, and they perform poorly on imbalance datasets. To solve these problems, this research proposes a hybrid model that combines the strengths of random forest and principal component analysis (PCA). The random forest model outperforms the linear regression model in the analysis, with a mean accuracy of 0.9730 versus 0.8695. With 5 best features, Random Forest Regression with Chi-Square feature selection had the best accuracy (0.9754 vs. 0.8663 for Linear Regression).

However, this study chose random forest as its evaluation method without providing any reason for why it did not investigate alternative empirically proven high-performing esemble methods such as bagging, boosting, Bayesian averaging, and stacking. The research might have used powerful algorithms like MLP and XGBoost, which have the ability to address the problem of Over-fitting as well as find relevant features using Variable Importance Operation. Again, when it comes to choosing on the cutoff points, PCA implementation allows for the modeler's subjectivity, suffocating statistical truth and independence. Furthermore, other evaluation metrics such as the R2, RMSE, were not used to verify the claims/results in this study.

Sharma et al. (2021) says that the main goal of their research work is to present supervised machine learning algorithms for predicting diamond prices (in US dollars). Due of their monetary value, precious stones such as diamonds are always in high demand. The cost of such stones varies depending on their characteristics. As a result, the study conducted a comparative analysis and application of multiple supervised models in predicting the diamond price. Because these heavy stones are more expensive than lighter stones, the relationship will not be linear.

This study compares and contrasts eight alternative supervised models, including linear regression, lasso regression, ridge regression, decision tree, random forest, ElasticNet, AdaBoost Regressor, and Gradient-Boosting Regressor, to find the best model for the job. According to the research given in the publication, the random forest method outperforms

the other supervised learning algorithms. The Random forest method can reach an R^2 score of 97.93% when the dataset is split 80% for training and 20 percent for testing, according to the paper.

When compared to MLR, which does not use the shrinkage idea, coefficient shrinkage models (lasso,ridge,elastic net) may produce overly optimistic findings. Random forest ought to have been compared to other novel machine learning algorithms such as XGBoost, MLP, BRT, SVR, and others, since the focus was on supervised model evaluation.

This study appears to dismiss the importance of classification in pricing (Similar studies have found that cut is an important variable in diamond pricing thus classifying stones by this variable would have been more informative). The models were not assessed using multiple regression metrics such as the MAE and R squared to further determine the claims/results.

Mihir et al. (2021) observes that unlike gold and silver, establishing the price of a diamond is extremely difficult since numerous factors must be taken into account, such as clarity, carat weight, cut, breadth, length, color, percentage of depth, and table width. The goal of this project is to develop the most efficient algorithm for predicting diamond prices.

Linear regression, Support Vector regression, Decision trees, Random Forest regression, kNeighbors regression, CatBoost regression, Huber regression, Extra tree regression, Passive Aggressive regression, Bayesian Regression, and XGBoost Regression are some of the algorithms used to train machine learning models on the diamond dataset for predicting diamond prices based on various attributes. CatBoost Regression was found to be the most suited algorithm for diamond price prediction, with the greatest R^2 score of 0.9872 and comparatively lower RMSE and MAE values, based on the performance parameter values and analysis.

To acquire more accurate findings, one of the future prospects of this article is to introduce a variety of factors such as shape, table value, polish, symmetry, and so on.

Chu (1996) suggests that the costs of diamond rings be related to the weights of their diamond stones using basic linear regression. Simple linear regression was employed to conduct the analysis in this study. The generated regression line has a negative intercept. The postulated

pricing mechanism implies a negative relationship between diamond ring costs and the weights of their diamond stones, raising doubts about the method's validity.

There is plenty of evidence that diamond essential characteristics have a non-linear connection, implying that the model utilized was incorrect. Furthermore, the technique relied solely on carat weight for price, with no explanation as to why other characteristics as as color, clarity, and cut were overlooked.

Chu (2001) attempts to build a diamond stone pricing model. The paper also teaches us about the different degrees of clarity and color, as well as the relative cost of cartage. The 4 C's: Carat, Clarity, Color, and Cut are elements that determine the price of a diamond stone, according to this research. Carat units are used to measure the weight of a diamond stone. A carat is the same as 0.2 grams. In the absence of other factors, bigger diamond stones attract greater prices due to their scarcity.

The pricing model is built using multiple linear regression (MLR), which is believed to provide flexibility and clarity when dealing with exogenous elements. Carat, color, clarity, and GIA and IGI certificates were the criteria utilized to predict the price. The value of r-squared was 97.2%.

According to the study, there is a non-linear relationship between caratage and price, with heavier stones being more valuable than lighter ones. A scatter plot of Price against Carats confirms this, with the trend appearing to fan out. As a result, instead of using MLR to construct a diamond pricing model, it would be more prudent to use machine learning models. Furthermore, other significant variables were kept out of the analysis.

Cardoso and Chambel (2005) propose new pricing models for cut diamonds. Derived models may have some advantages over the traditional Rapaport - an industry-wide adopted price indicator - in that they are based on published final selling prices, which already include corrections not included in the Rapaport, and so can evaluate prices closer to the market.

This is accomplished through the use of regression trees, Chi-Square Automatic Interaction Detection, and neural networks (with backpropagation). Neural networks outperform other

methods in terms of prediction, accounting for almost 96 percent of the fluctuation in cut diamond unit pricing.

The research did not take into account novel ensemble techniques used in machine learning, such as Random Forest, XGBoost, and others. The researcher ought to have used a Multi-layer perceptron to compare the results to a single hidden layer perceptron. The study does not rigorously assess the proposed prediction model's flaws, such as the RMSE, neglecting a crucial statistical decision-making method.

Scott and Yelowitz (2010) takes diamond into account when examining the market for commodities that are consumed not just for their intrinsic utility but also for the impact their usage has on others. Diamonds are in high demand because they create a market for social status and the inherent usefulness that comes with wearing beautiful things. Data was gathered from online diamond sellers in order to investigate the factors of diamond prices empirically. Carat weight and cut are established during the production process, whereas color and clarity are dictated by nature.

The first standard considers carat weight, color, cut, and clarity when determining the log of price. For Blue Nile, Union Diamond, and Amazon listed diamonds, this results in adjusted R squared values of 88.9%, 89.8%, and 93.7%, respectively. All round diamonds between 0.4 and 0.6 carats are included in the sample.

Given the non-linear nature of the relationship between diamond attributes and price, this research should have looked into using machine learning to solve the problem. Furthermore, the study does not use error measurements such the RMSE to corroborate the findings.

Chapter 3

Methodology

3.1 Introduction

This chapter gives detailed information and description of research methodology to be used. We propose SML models as outlined in Figure 3.1 and in Figure 3.2.

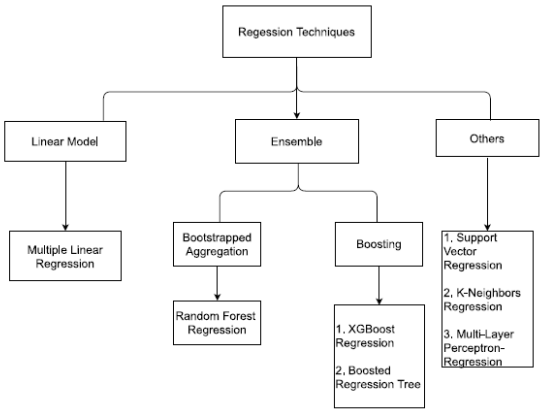

Figure 3.1: Regression Techniques

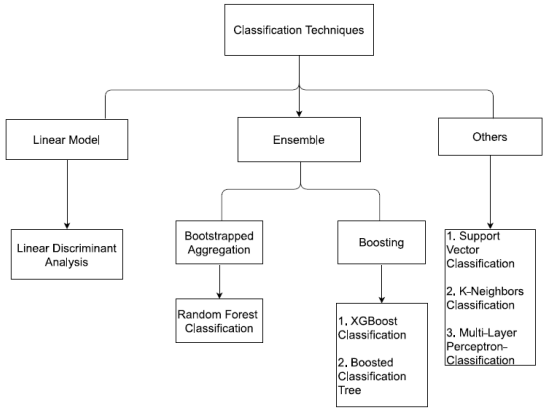

Figure 3.2: Classification Techniques

3.2 Multiple Linear Regression (MLR)

We consider this model when the study variable involves more than one predictor variables. Here, the relationship is important in that it allows the mean function $E(y)$ to depend on more than one predictor variables and to assume shapes other than straight line (Montgomery and Runger, 2010).

Given the model as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \tag{3.1}$$

now, given that the $n-tuples$ of observations follow the same model, below is satisfied:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_k X_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_k X_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_k X_{nk} + \varepsilon_n \end{aligned} \tag{3.2}$$

The above n equations can be expressed in form of matrices as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}}_{DesignMatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3.3}$$

Through algebraic operation, the OLS estimator of β is given as:

$$\beta = (X'X)^{-1}X'y \tag{3.4}$$

3.3 Boosted Classification and Regression Trees (BCARTs)

Tree boosting is a method of combining many weak learners (trees) into a strong classifier where: Each tree is created iteratively and the tree's output $h(x)$ is given a weight w relative to its accuracy.

The ensemble output is the weighted sum:

$$\hat{y}(x) = \sum_t w_t h_t(x) \tag{3.5}$$

After each iteration each data sample is given a weight based on its misclassification i.e. the more often a data sample is misclassified, the more important it becomes. Here, the goal is to minimize an objective function:

$$O(x) = \sum_i l(\hat{y}_i, y_i) + \sum_t \Omega(f_t) \tag{3.6}$$

where:

- $l(\hat{y}_i, y_i)$ is the loss function i.e. the distance between the truth and the prediction of the ith sample.
- $\Omega(f_t)$ is the regularization function i.e. it penalizes the complexity of the tth tree.

Lampa et al. (2014) observes that CARTs are incredibly straightforward yet effective. They divide the data into a number of isolated zones and, within each of these parts, approximation the result with a constant value. A sequence of binary splits in the input variables are used to achieve this. A statistical criterion, such as the residual sum of squares, is optimized by first identifying the variable and split point that best fits the CART.

Utilizing the subset of observations that passed through the preceding split, the optimal split is found within each generated subset. This is done repeatedly until there are normally no more than 10 observations left that can be split. Single CARTs are referred to as weak learners in statistical learning terminology because of their inferior prediction abilities. The concept behind stochastic gradient boosting (boosting), a numerical approach, is that a strong learner with improved prediction performance can be generated by combining several weak learners.

Using a function $F(x)$, commonly referred to as the target function and approximated via an additive expansion, the objective is to accurately map a set of explanatory variables x to an outcome variable y.

Selection bias in favor of variables with a large number of potential split points is a downside. CARTs' high degree of variability is another problem; even a little change in the outcome data can result in a different CART.

$$\hat{F}(x) = \sum_{m=1}^{M} \beta_m b(\mathbf{x}; \gamma m)$$

.

Where M is the number of weak learners; βm are the expansion coefficients and b($\mathbf{x}; \gamma m$)are individual weak learners characterized by the parameters γm. Accuracy is defined by a loss function L(y, F) which represent the loss in predicting y with F($\mathbf{x}$).

The algorithm works as follows;

1. Initialize $\hat{F}_0(\mathbf{x})$ to a constant α.
2. Randomly sample a fraction η from the data without replacement.
3. Using η, compute the negative gradient of the loss function, $z_m = -\nabla L$, and fit a depth d CART, g($\mathbf{x}$), predicting z_m.
4. Update $\hat{F}_m(\mathbf{x}) \leftarrow \hat{F}_{m-1}(\mathbf{x}) + \lambda \rho g(\mathbf{x})$.
5. Iterate steps 2 through 4 M times.

.

In step 4, ρ is the step size along the gradient and is a shrinkage parameter which slows down the learning to reduce overfitting. The parameters M, d and λ can be tuned using the bootstrap or cross-validation.

Further details on BCARTs can be obtained from Friedman (2002) and (Breiman et al., 1984).

3.4 eXtreme Gradient Boosting (XGBoost)

The XGBoost algorithm tries to minimize the following objective function (loss function and regularization) J at step t:

$$J^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{t-1} + f_t(x_1)\right) + \sum_{i=1}^{t} \Omega(f_i) \tag{3.7}$$

where the first term contains the train loss function L (e.g. mean squared error) between real class y and output $\hat{y}$ for the n samples and the second term is the regularization term, which controls the the complexity of the model and helps to avoid overfitting (Dimitrakopoulos et al., 2018). It is observable that the XGBoost objective is a function of functions (i.e. L is a function of CART learners, a sum of the current and previous additive trees). To solve the above objective function, Taylor approximation is applied to transform the original objective function to a function in the Euclidean domain, in order to be able to use traditional optimization techniques. In XGBoost, the complexity is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3.8}$$

where T is the number of leaves, γ is the pseudo-regularization hyperparameter, depending on each dataset and λ is the $L2$ norm for leaf weights.

Using gradients for second-order Taylor approximation of the loss function and finding the optimal weights w, the optimal value of objective function is:

$$J(t) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} + \gamma T \tag{3.9}$$

where $g_i = \partial_{\hat{y}^{t-1}} L(y, \hat{y}^{t-1})$ and $h_i = \partial^2_{\hat{y}^{t-1}} L(y, \hat{y}^{t-1})$ are the gradient statistics on the loss function, and I is the set of leaves.

The XGBoost benefits from the shrinkage strategy in which newly added weights are scaled after every step of boosting (greedy algorithm) by a learning factor rate. This helps to diminish the effects of future new trees on every existing individual tree, thereby reducing the risk of overfitting (Mohammadi et al., 2021).

XGBoost is comprised of three main elements:

- **Weak Learners** – simple decision trees that are constructed based on purity scores (e.g., Gini)
- **Loss Function** – a differentiable function you want to minimize. In regression, this could be a *mean squared error*, and in classification, it could be *log loss*.
- **Additive Models** – additional trees are added where needed, and a functional gradient descent procedure is used to minimize the loss when adding trees.

3.5 Support Vector Machine (SVM)

SVM is a machine learning technique that works by identifying the optimal decision boundary that separates data points from different classes, and then predicts the class of new observations based on the said boundary. Kassambara (2017) observes that SVM can be used for two-class as well as multi-class classification problems.

James et al. (2013) asserts that there is an extension of the SVM for regression (i.e. for a quantitative rather than a qualitative response), called *support vector regression*. Support vector regression seeks coefficients $(\beta_o, \beta_1, \ldots, \beta_p)$ that minimize a different type of loss, where only residuals larger in absolute value than some positive constant contribute to the loss function.

Suppose that we have $n \times p$ matrix of data set, where samples belong to two linearly separable classes represented by $+1$ or 1, and suppose g_i is the features vector. The, $(g_i, y_i) \in G \times Y, i = 1, 2, \ldots, n$ will be satisfied where $y_i \in (+1, -1)$ is the target variable dichotomy in the p dimensional space. The aim is to classify the sample into one of the two classes and by extension find an SVM classifier that generalizes to a multi-class problem achieved by finding an optimal separating hyperplane (Mohammed et al., 2021).

A separating hyperplane for the two classes is given as:

$w * g + b \geq 1$ when $y_i = +1$.

$w * g + b \leq -1$ when $y_i = -1$.

where w is the weight vector, b is the bias, and $|b|/||w||$ is the perpendicular distance to the hyperplane. The distance from the nearest point in each class to the hyperplane becomes $1/||w||$ and $2/||w||$ between the two classes after rescaling. The solution to the optimization problem is obtained by maximizing the margin:

$min_{w,b}||w||^2$

subject to $y_i(w*g+b) \geq 1, i = 1,2,\ldots,n.$

In this study, we will employ one-vs-one multi-class classification in which the SVM classifier produces all possible pairs of binary classifications. Here, given that we have k classes where $k > 2$, it follows that $\frac{k(k-1)}{2}$ binary classifiers are produced in the training step of the algorithm. Consequently, a sample in the test dataset is assigned the class label that is voted the most by the binary classifiers from the trained *one-vs-one* SVM.

3.6 K-Nearest Neighbors (KNN)

K-nearest neighbors (kNN) is a non-parametric method used for classification and regression (Yao and Ruzzo, 2006). Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by ψ_0. It then estimates the conditional probability for class j as the fraction of points in ψ_0 whose response values equal j:

$$P_r(Y = j|X = x_0) = \frac{1}{K}\sum_{i\in\psi_0} I(y_i = j). \tag{3.10}$$

Lastly, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability (James et al., 2013).

The regression seeks to estimate $f(x_0)$ using the average of all the training responses in ψ_0, mathematically expressed as:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \psi_0} y_i. \tag{3.11}$$

3.7 Random Forests (RFs)

Random Forest is an unpruned classification or regression tree ensemble produced by employing bootstrap samples of the training data and random feature selection in tree induction. The ensemble's forecasts are summed (majority vote or averaging) to make a prediction. When creating these decision trees, a random sample of m predictors is picked as split candidates from the entire set of p predictors each time a split in the tree is evaluated. Only one of the m predictors can be used in the split. A fresh sample of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$, i.e. the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (James et al., 2013).

The random forest prediction is the most prevalent class among individual tree predictions in the *classification* setting. If there are T trees in the forest, the amount of votes a class m receives is:

$$v_m = \sum_{t=1}^{T} I(\hat{y}_t == m). \tag{3.12}$$

where $\hat{y}_t$ is the prediction of the $t-th$ tree on a particular instance. The indicator function $I(\hat{y}_t == m)$ takes on the value 1 if the condition is met, else it is 0.

In a regression setting, the random forest's forecast is the average of the individual trees' predictions. If there are T trees in the forest, each making a prediction $\hat{y}_t$, the final prediction $\hat{y}_t$ is:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t. \tag{3.13}$$

3.8 Multi-Layer Perceptron (MLP)

MLP is an algorithm that is inspired by the structure and function of the brain, which is usually called Artificial Neural Networks (ANN). To store information, the brain changes the connections between neurons. The neuron does not store information; instead, it enables signal transmission between neurons. The human brain is made up of a gigantic network of neurons. The neural network mimics the brain's mechanism. While the human brain employs neuronal association, the neural network employs neuronal connection weights (Ghatak, 2019). The information of the neural network is stored in the form of weights and biases as demonstrated in Figure 3.3.

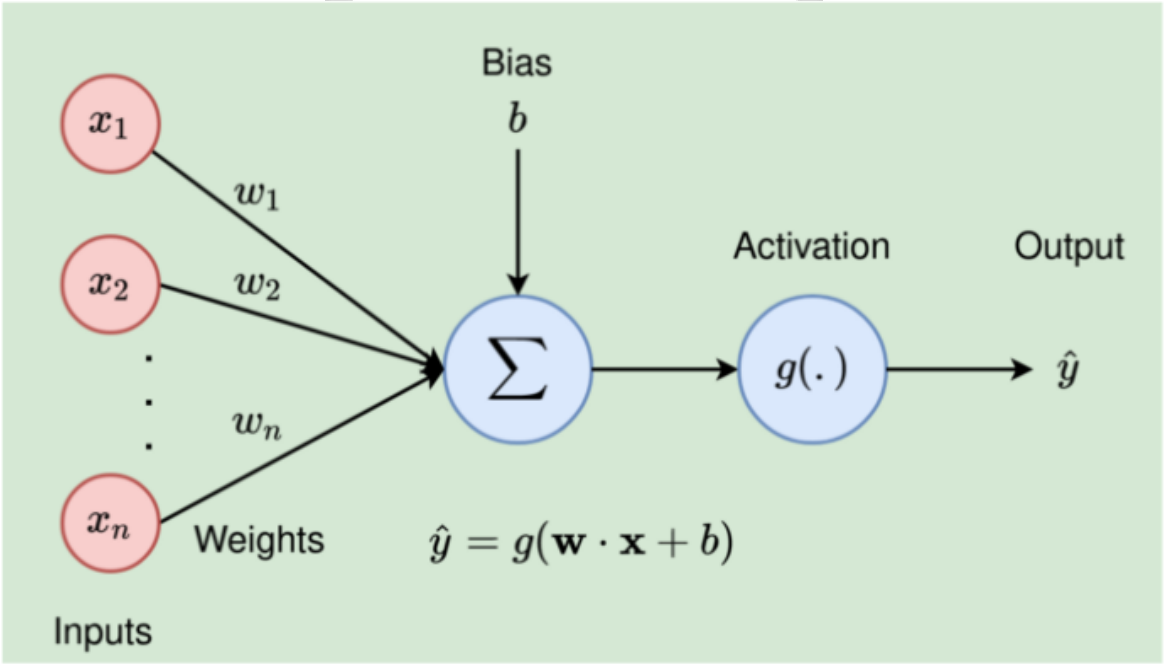

Figure 3.3: Neural Network Architecture

The input signals are multiplied by the weights before entering the node as shown below

$$v = (w_1 \times x_1) + (w_2 \times x_2) + (w_3 \times x_3) + b = \mathbf{wx} + b \tag{3.14}$$

The weighted sum can be expressed in matrix form

$$\mathbf{v} = \begin{bmatrix} w_1 & w_2 \dots w_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b \end{bmatrix} \tag{3.15}$$

The output of the node (y) is processed using *activation function* (g) as shown below

$$\hat{y} = g(v) = g(\mathbf{w}.\mathbf{x} + b) \tag{3.16}$$

It is important to note that MLP is defined by two or more hidden layers. According to Cardoso and Chambel (2005), the more hidden units there are in a network, the less likely it is to encounter a local minimum during training. Figure 3.4 shows a typical MLP network.

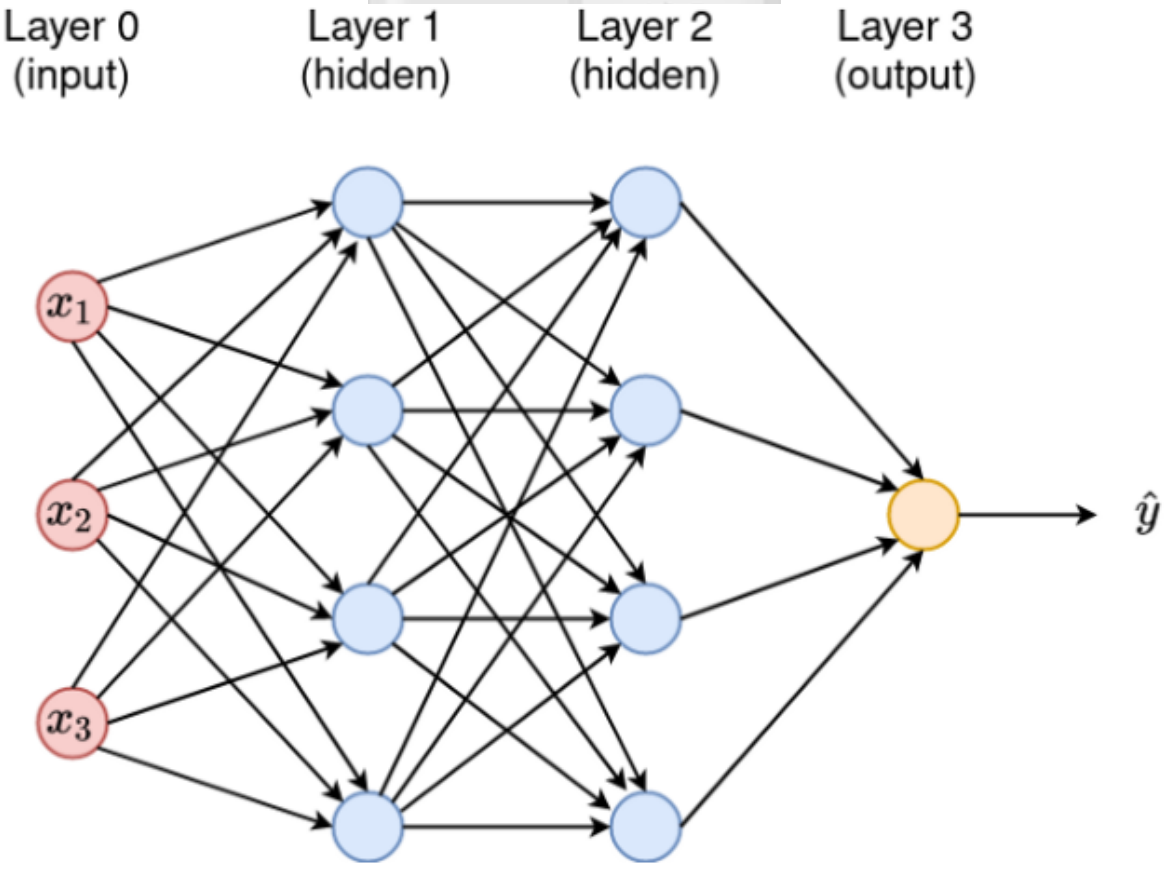

Figure 3.4: Multi-Layer Perceptron Architecture

Input nodes merely relay the input signal; they do not compute the weighted sum or apply the activation function. Because they are not visible from outside the neural network, they are called hidden layers. In supervised learning, the learning rule trains the neural network to produce the proper output that has already been determined.

The weights are initialized and the error is calculated accordingly. Then, the weights are adjusted to reduce the error. This procedure is repeated until the minimum error is attained. The systematic way of modifying the weights is known as the *Learning Rule*, as demonstrated by the *generalized delta rule* below

$$\Delta wij \leftarrow w_{ij} + \alpha \delta_i x_j$$

Where;

- $\delta_i = \phi_i'(v_i)e_i$
- e_i is the error of node i
- $\underbrace{e_i}_{error} = \underbrace{d_i}_{correct\ out put} - \underbrace{y_i}_{observed\ out put}$
- $\phi' = \frac{d}{dx}$ (*Activation Function*)
- Δw_{ij} is the updated weight
- w_{ij} is the previous weight
- x_j is the output from node j ($j = 1, 2, 3 \ldots$)
- α is the learning rate ($0 \leq \alpha \leq 1$)

The learning rate determines the extent to which weights are changed at every epoch. A high value of α indicates that the output gravitates around the expected solution while a low value shows that out fails to converge to the acceptable solution. Sigmoid, given as below, will be used as the activation function.

$$\phi(X) = \frac{1}{1+e^{-X}} \tag{3.17}$$

Getting the first derivative;

$$\phi'(x) = \phi(x)(1-\phi(x)) \tag{3.18}$$

The strengths of Multi-Layer Perceptron (MLP);

- The impediment of training multi-layers is solved by Back-propagation algorithm.
- Poor performance due to vanishing gradient is addressed by use of *Rectified Linear Unit* (ReLU), which is applied as the activation function.
- The vulnerability to overfitting resulting from model complexity with additional hidden layers is solved by 'Dropout', i.e.training some of the randomly selected nodes rather than the entire network. Regularization is also used to prevent overfitting by simplifying the architecture of MLP.
- The *softmax* activation function in the output layer helps to keep range between 0 1 which can be used as probabilities.

ReLU function gives us the maximum value between zero and a given input.

$$\phi(x) = \begin{cases} x, x \geq 0 \\ 0, x \leq 0 \end{cases} = max(0,x) \tag{3.19}$$

The derivative of ReLU function;

$$\phi'(x) = \begin{cases} 1, x \geq 0 \\ 0, x \leq 0 \end{cases} \tag{3.20}$$

The softmax activation function;

$$\phi(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{3.21}$$

Where;

- $\phi = softmax$
- $\mathbf{z} = imput\ vector$
- $e^{z_i} = standard\ exponential\ function\ for\ input\ vector$
- $K = number\ of\ classes\ in\ the\ multi-class\ classifier$
- $e^{z_j} = standard\ exponential\ function\ for\ output\ vector$

3.9 Linear Discriminant Analysis (LDA)

This extends the LDA classifier to the case of multiple predictors. Here, the assumption is that $X = (X_1, X_2, \ldots, X_p$ is drawn from a *multivariate normal* or *multivariate Gaussian* distribution $N(\mu k, \Sigma)$, with a class-specific multivariate mean vector and a common covariance matrix (James et al., 2013).

Chris (2021) postulates that LDA uses *Bayes Theorem* for classification which we can explain by noting that if we have K classes and we want to classify the qualitative response variable Y where there are K possible distinct and unordered values derived as follows:

Let $\pi_k$be the prior probability that a given randomly chosen observation comes from the kth class. Let $f_k(x) \equiv P_r(X = x|Y = k)$ be the density function of X for an observation from the kth class. $f_k(x)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$ and $f_k(x)$ is relatively small if it is very unlikely that an observation in the kth class has $X \approx x$.

Bayes Theorem states that:

$$P_r(Y = k|X = x) = \frac{\pi k f_k(x)}{\sum_{l=1}^{K} \pi l f l(x)} \tag{3.22}$$

Letting $p_k(x) = P_r(Y = k|X)$, we can simply plug in estimates of π_k and $f_k(X)$ into the formula which can be generated with the software that then takes care of the rest. We refer to $p_k(x)$ as the *posterior* probability that an observation $X = x$ belongs to the kth class given the predictor value for that observation.

Estimating $_k$ is easy if we have a random sample of $Y's$ from the population but estimating $f_k(X)$ is more difficult. However, if we have an estimate for $f_k(x)$ then we can build a classifier that approximates the Bayes classifier.

By assuming that $X = (X_1, X_2 \ldots, X_p)$ is drawn from a multivariate Gaussian distribution, with a class specific mean vector and a common covariance matrix which we can write as

$$X \sim N(\mu, \Sigma)$$

to indicate that p has a multivariate Gaussian distribution.

$$E(X) = \mu$$

is the mean of the X vector with p components and

$$Cov(X) = \Sigma$$

is the pp covariance matrix of X.

Formally, the multivariate Gaussian density is given as:

$$f(x) = \frac{1}{(2_\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \tag{3.23}$$

Plugging the density function for the kth class, $f_k(X = x)$ into equation 19 above, and applying some algebra we see that the Bayes classifier assigns $X = x$ to the class for which:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log\ \pi_k \tag{3.24}$$

is the largest. The Bayes decision boundaries represent the set of values x for which $\delta_k(x) = \delta_l(x)$. In other words for which $x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k = x^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l$, for $k \neq l$

The log $_k$ term has disappeared because each of the three classes has the same number of training observations, thus $_k$ is the same for each class. To estimate $\mu_1 \ldots, \mu_k, \pi_1 \ldots, k$ and Σ we use similar conventions for the case where $p = 1$

3.10 Regression Evaluation Metrics

Root Mean Squared Error (RMSE)

The root-mean-square error measures the model's prediction error. It is the average difference between the observed known values of the outcome and the predicted values by the model (Kassambara, 2018). Low values of RMSE signifies better predictions from the model. Barnston (1992) gives below mathematical expression for RMSE:

$$\text{RMSE}_{fo} = \left[\sum_{i=1}^{N} (z_{fi} - z_{oi})^2 / N \right]^{\frac{1}{2}} \tag{3.25}$$

Where;

- $f = forecasts\,(expected\,values\,or\,unknown\,results)$
- $o = observed\,values, (known\,results)$
- $(z_{fi} - z_{oi})^2 = differences, squared.$
- $N = sample\,size.$

Mean Absolute Error (MAE)

Willmott and Matsuura (2005) Assert that MAE as a measure of model's accuracy is unambiguous, stable and more natural measure of average error unlike RMSE which varies with variability within the distribution of error magnitudes. MAE is calculated by summing the magnitudes (absolute values) of the errors to obtain the 'total error' and then dividing the total error by n, as shown below;

$$\text{MAE} = \left[n^{-1} \sum_{i=1}^{n} |e_i| \right] \tag{3.26}$$

Where;

$P_i = 1,2,\ldots,n$, $O_i = 1,2,\ldots,n$, $e_i = 1,2,\ldots,n$ and

- $e_i = P_i - O_i$
- $n = sample\,size$
- $P = predicted\,values$
- $O = observed\,values$

Low values of MAE signifies better model performance in terms of the accuracy of predictions.

R Squared

The R-squared (R^2) is given by a series of some metrics including;

residual sum of squares (RSS) expressed as:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{3.27}$$

residual standard error (RSE) given as:

$$\text{RSE} = \sqrt{\frac{1}{n-2} \times \text{RSS}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.28}$$

where y_i and $\hat{y}_i$ are the actual and predicted values of observations, respectively.

total sum of squares (TSS) which is the total variance in the response Y i.e.inherent in the response before the regression is performed (James et al., 2013). It is given by the formula;

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.29}$$

where $\bar{y}$ is the mean of observed response values.

The R-squared (R^2) statistic,commonly referred to as coefficient of determination is thus given as

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{3.30}$$

While $TSS - RSS$ measures the amount of variability in the response that is explained (or removed) by performing the regression, R^2 is a measure of the proportion explained variability in the response variable Y that is associated with predictor variable X.

Wooldridge (1991) propose the following formula for the adjusted R^2:

$$\bar{R}^2 \equiv 1 - \frac{RSS/(T-K-1)}{TSS/(T-1)} \tag{3.31}$$

Where;

- $T = sample\,size$
- $K = number\,of\,predictors$

The adjusted R^2 is modified or adjusted so as to accommodate the changes in degrees of freedom that results due to addition or removal of some independent variables in a regression model.

3.11 Classification Evaluation Metrics

Confusion Matrix

Table 3.1 is a tabular representation of *Actual* vs *Predicted* values. It helps us find the accuracy of the model thus avoid overfitting. Accuracy refers to the total number of predictions that were correct. The below table summarizes the elements of the Confusion Matrix.

	Predicted(Y)	Predicted(Y)
Positive(Y=1)	True Positive(A)	False Negative
Negative(Y=0)	False Positive(C)	True Negative(D)

Table 3.1: Confusion Matrix Table

$$\text{Accuracy} = \frac{A+D}{A+B+C+D} \tag{3.32}$$

In essence, this is given as

$$\frac{TP+TN}{TP+FP+TN+FN}$$

Proportion of the predicted positive cases that were correct, or *Precision* is given by:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.33}$$

Proportion of the positive cases that were correctly identified, or *Recall* is given by:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3.34}$$

The harmonic mean of *Precision* and *Recall*, or *F-measure* Vafeiadis et al. (2015) is given by:

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.35}$$

Since good performance of a classifier cannot be exclusively measured by either precision or recall, F1-Score, which combines the two is used as a single metric for evaluating a classifier's performance (Vafeiadis et al., 2015). A F-measure value closer to one indicates better classifier performance.

The cohen's kappa measure

Cohen's kappa is a measure of the agreement between two raters who each classify N items into C mutually exclusive categories. The definition of k is:

$$k = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{3.36}$$

p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement.

$$p_o = \frac{A + B}{A + B + C + D} \tag{3.37}$$

$$p_{pos} = \frac{a + b}{a + b + c + d} \times \frac{a + c}{a + b + c + d} \tag{3.38}$$

$$p_{neg} = \frac{b + d}{a + b + c + d} \times \frac{c + d}{a + b + c + d} \tag{3.39}$$

$$p_e = p_{pos} + p_{neg} \tag{3.40}$$

The flowchart below illustrates the overall modeling process:

3.12 Overall Modeling Process

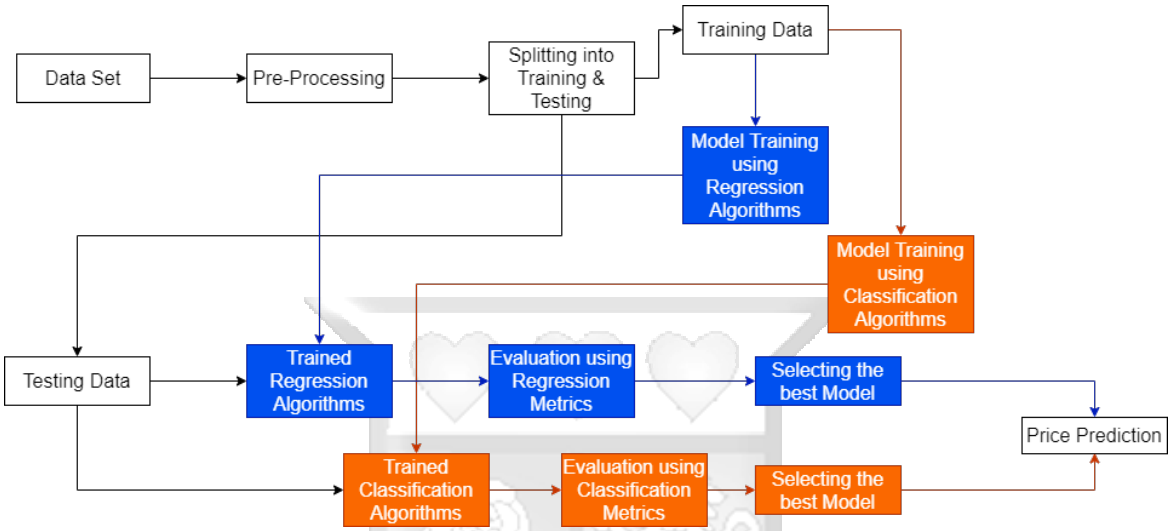

Figure 3.5: Overall Modeling Process

Figure 3.5 depicts a flow chart showing the various options for achieving the final goal, which is diamond price forecast. While the blue coloring represents an existing technique to diamond pricing that uses machine learning, the orange shading represents the study's recommended solution.

3.13 Data Type and Source

To understand and validate the case under inquiry, we looked at two sets of data. The first data set was created using simulation in order to mimic the real data in Chapter 4. There were 3000 observations and 7 variables in this dataset. The renowned iris flower data set, often known as Fisher's Iris data set, is a multivariate data set created by British statistician and biologist Ronald Fisher. (Fisher, 1936).

3.14 Simulated Data Analysis

Simulating a categorical variable

Let's say that I have a categorical variable (Group) which can take the values A, B, C and D. The aim is to generate 3000 random data points and control for the frequency of each as below:

$A = 20\%, B = 25\%, C = 25\%, D = 30\%$

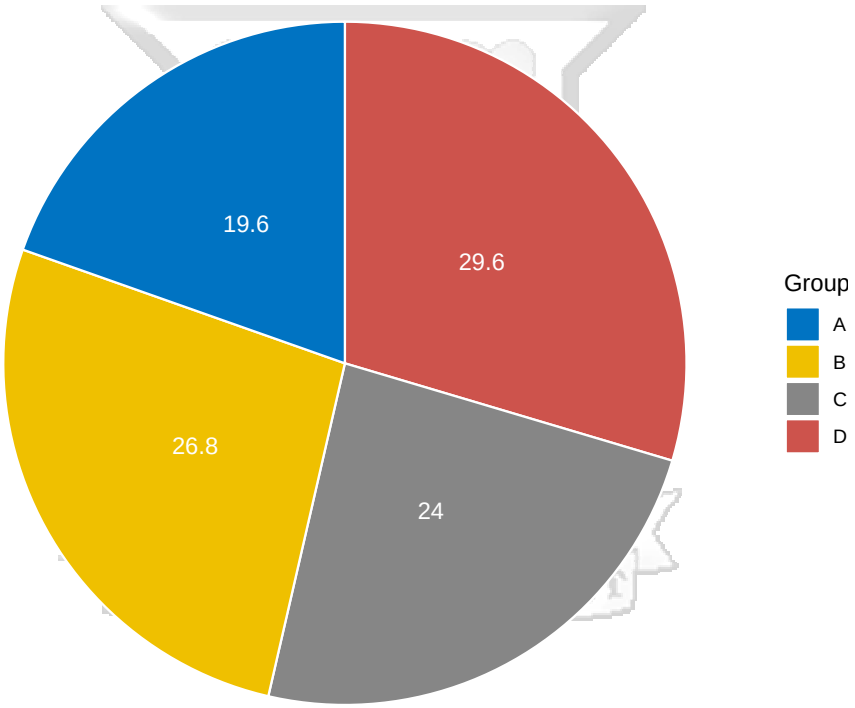

Figure 3.6: The Simulated Group Proportions

Figure 3.6 visualizes the simulated *Categorical* variable proportions.

The graphical features of simulated numeric variables

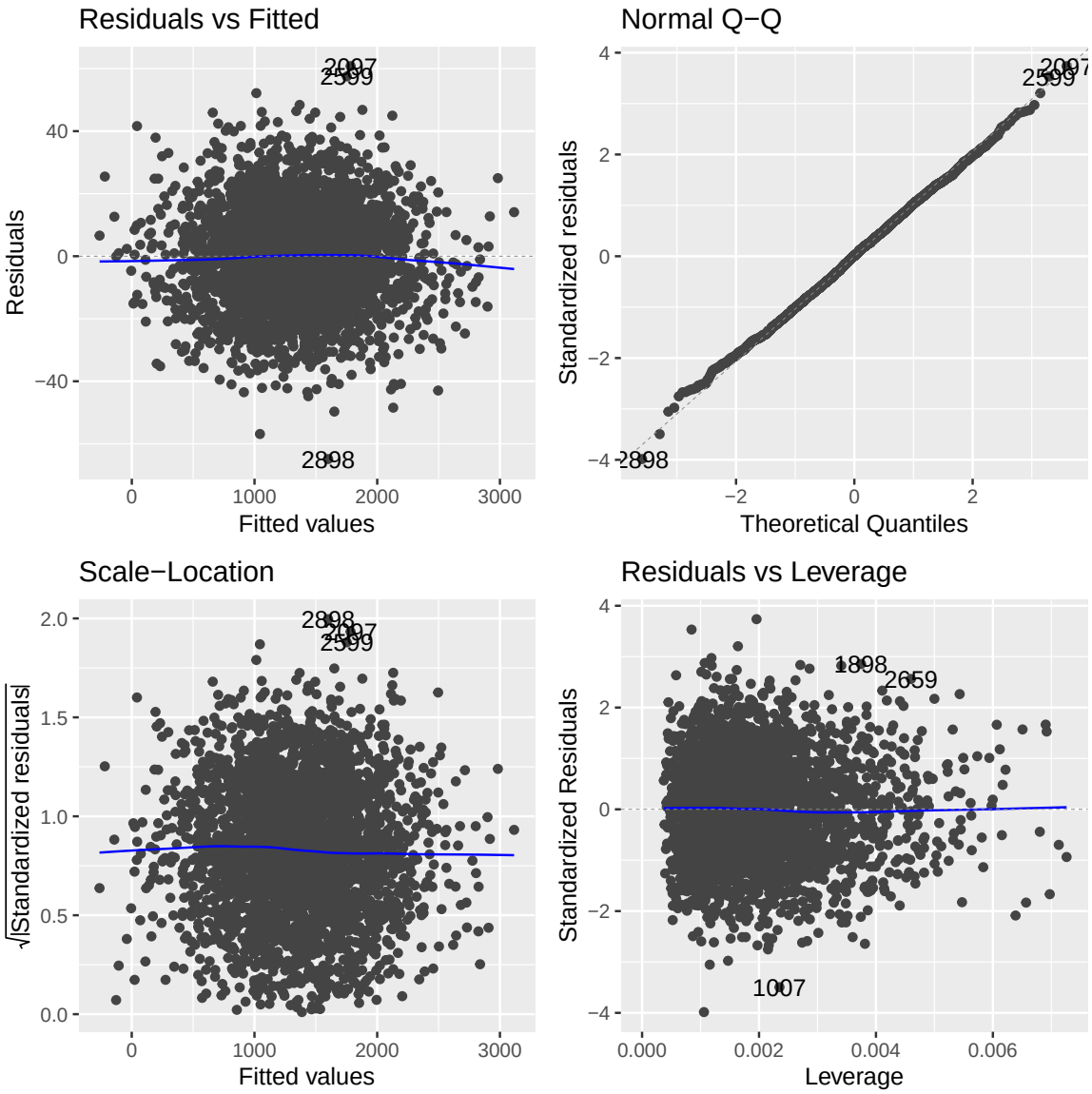

Figure 3.7: Multiple Linear Regression Assumptions

Figure 3.7 confirms that the simulated multivariate data set satisfy the following assumptions:

- Fixed location (Measure of central tendancy)
- Fixed scale (Measure of spread)
- Fixed distribution (Multivariate Normality)
- Randomness.

Table 3.2: Glimpse of the simulated data for 8 random observations

Group	Y	X1	X2	X3	X4	X5
D	1604.96	66.30	112.65	328.96	168.79	205.08
D	1303.96	49.11	229.80	165.10	107.73	105.87
B	2360.40	46.40	345.55	354.48	113.03	300.47
C	1267.44	46.40	146.64	223.07	67.73	186.52
A	1207.57	39.49	149.98	212.38	63.95	108.43
D	1943.86	54.20	187.04	409.62	84.00	188.72
C	1181.76	48.47	240.18	102.82	162.11	106.20
A	386.79	68.52	122.98	35.90	61.63	22.97

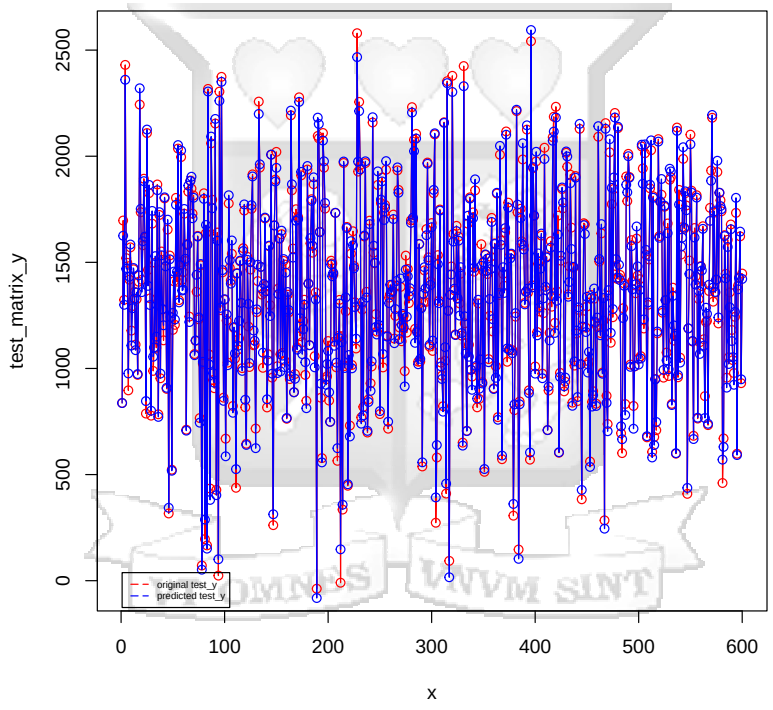

Figure 3.8: XGBoost Predicted Vs Actual

Figure 3.8 visually depicts strong convergence of predicted and actual data points for the XGBoost model's test set.

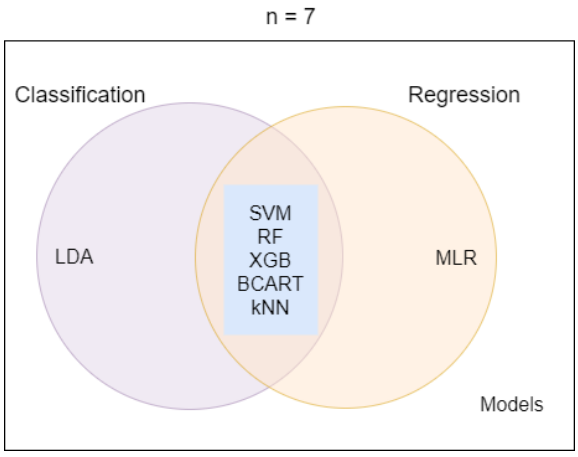

Figure 3.9: The Models

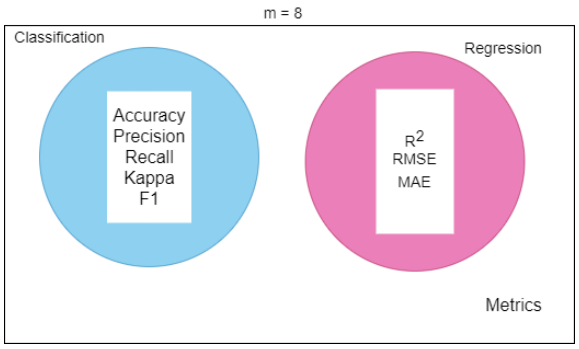

Figure 3.10: The Metrics

3.15 Simulation Analysis Results

Table 3.3: Regression Evaluation Metrics

Algorithms	R^2	RMSE	MAE
MLR	99.89	16.11	12.95
SVR	99.86	18.29	14.77
BRT	99.56	30.62	23.96
XGBoost	99.33	37.95	29.70
Rf	98.40	63.45	46.10
kNN	93.15	141.67	103.64

Table 3.4: Classification Evaluation Metrics

Algorithms	Precision	Recall	F1	Accuracy
XGBoost	26.84	24.91	18.57	31.17
SVM	13.33	25.06	12.54	29.60
kNN	26.56	26.96	25.85	29.10
LDA	14.24	24.54	15.79	28.6
Rf	23.62	23.86	23.15	25.08
BCAT	22.19	22.10	20.96	24.25

The regression results from simulated dataset are shown in Table 3.3 while classification results are presented in Table 3.4.

We will not explore MLR and LDA based on the analyses in Figures 3.9 and 3.10 because they do not apply to both scenarios under consideration, i.e. Regression and Classification.

$Overall\ Performance = \frac{X}{7} * 100$

X=Total metrics under investigation.

Algorithms	R^2	RMSE	MAE	Precision	Recall	F1	Accuracy
SVM	1	2	2	5	2	5	2
BCART	3	3	3	4	5	3	5
XGBoost	4	4	4	1	3	4	1
Rf	4	4	4	3	4	2	4
kNN	5	5	5	2	1	1	3

Table 3.5: The Lead Table

Algorithms	Rating (%)
SVM	14.29
BCART	0.00
XGBoost	28.57
Rf	0.00
kNN	14.29

Table 3.6: Overall Algorithm's Performance

Table 3.5 shows the algorithms' ranking based on evaluation metrics in both regression and classification. Table 3.6 presents the overall algorithms' performance rating in percentage where the best ML model in regression and classification is *XGBoost* at 28.57%.

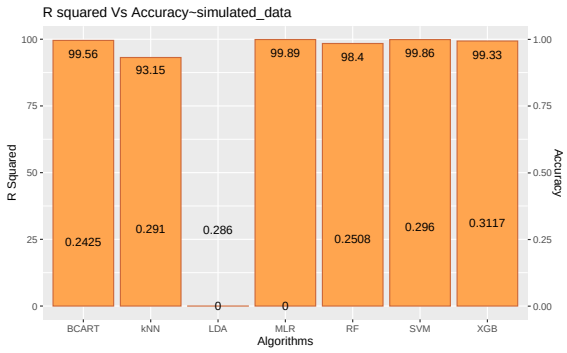

Figure 3.11: R squared Vs Accuracy

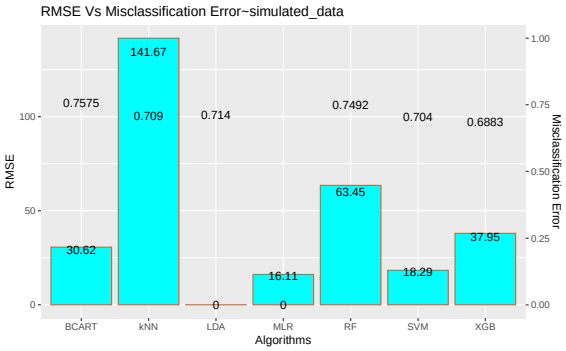

Figure 3.12: RMSE Vs Misclassification Error

The results in Figure 3.11 rate SVR (svmLinear) and BRT models highest for regression at R^2 values of 99.86% and 99.56% respectively . XGBoost is the best for classification at an accuracy value of 31.17%. kNN is the weakest in regression at R^2 value of 93.15%. Overally, XGBoost shows better results for predicting categorical and numeric response variables.

3.16 iris Data Analysis Results

Table 3.7: Regression Evaluation Metrics

Algorithms	R^2	RMSE	MAE
MLP	93.3	0.24	0.14
MLR	88.77	0.32	0.26
Rf	86.62	0.28	0.23
BRT	86.04	0.30	0.24
XGBoost	81.28	0.40	0.31
SVR	80.96	0.34	0.28
kNN	80.44	0.36	0.32

Table 3.8: Classification Evaluation Metrics

Algorithms	Precision	Recall	F1	Accuracy
LDA	100.00	100.00	100.00	100.00
kNN	96.97	96.67	96.66	96.67
Rf	94.44	93.33	93.27	93.33
SVM	93.33	93.33	93.33	93.33
XGBoost	93.33	93.33	93.33	93.33
MLP	88.89	94.87	90.56	92.31
BCT	92.31	90.00	89.77	90.00

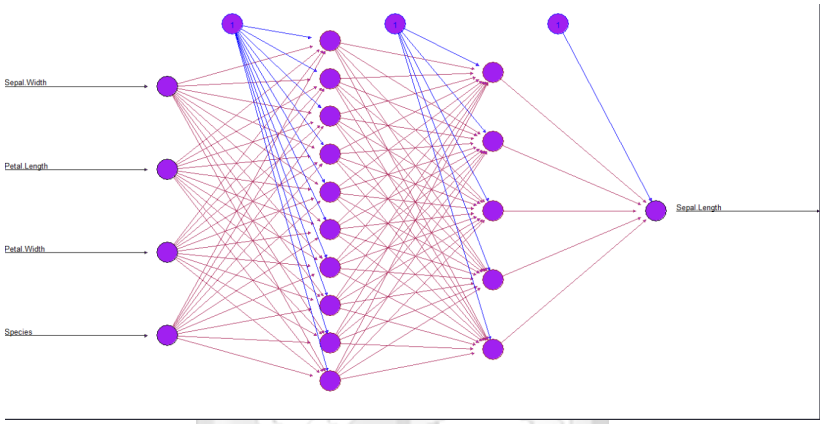

Figure 3.13: Multi-Layer Perceptron Architecture

Table 3.9: Algorithm's Lead Table

Algorithms	R^2	RMSE	MAE	Precision	Recall	F1	Accuracy
SVM	4	3	3	3	3	3	3
BCART	2	2	2	5	5	5	5
XGBoost	3	5	4	4	4	4	4
Rf	2	2	2	2	2	3	2
kNN	5	4	5	1	1	1	1
MLP	1	1	1	6	2	5	5

Table 3.10: Classification Evaluation Metrics

Algorithms	Rating (%)
SVM	0.00
BCART	0.00
XGBoost	0.00
Rf	0.00
kNN	57.14
MLP	42.86

$$Overall\ Performance = \frac{X}{7} * 100$$

Table 3.10 shows that MLP is the top ML model for regression at 42.86%, while kNN leads in classification metrics at 57.14%. The MLP model is made up of two hidden layers, each with 10 and 5 neurons as shown in Figure 3.13

Figure 3.14 demonstrates that kNN has the best classification accuracy of 96.97%, while MLP has the best regression results with a R^2 value of 92.31%. Despite the fact that MLP

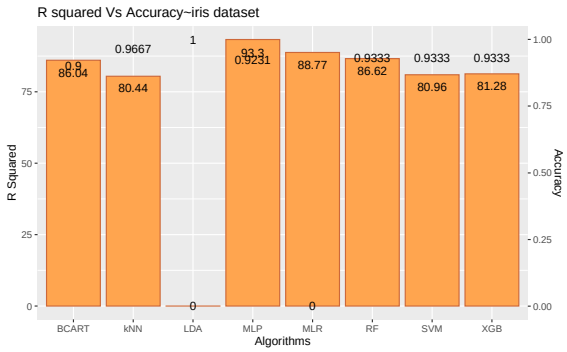

Figure 3.14: R squared Vs Accuracy

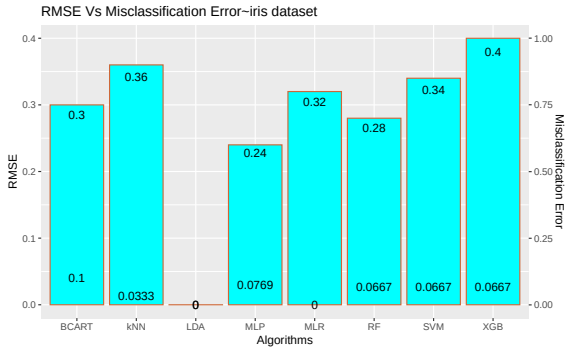

Figure 3.15: RMSE Vs Misclassification Error

produces great results in regression metrics, the study will not use it due to the high computing cost of parameter tuning and training time.

Chapter 4

Data Analysis

4.1 Introduction

The focus of this chapter is data analysis and interpretation of results. It entails exploratory analysis, descriptive statistics and correlation analysis to estimate the relationship between variables. The entire analysis was carried out in *R*.

4.2 Data Type and Source

Kaggle, a data repository with thousands of datasets, was used in the investigation. It is an online community for machine learning practitioners and data scientists, as well as a robust, well-researched, and sufficient resource for analyzing various data sources. On Kaggle, users can search for and publish various datasets. In a web-based data-science environment, they can study datasets and construct models. The Diamond Dataset's primary features will be provided by the Kaggle Diamond Dataset, which has approximately 53000 observations. (Agrawal, 2017).

4.3 Exploratory Data Analysis

This is a way of evaluating data sets in order to summarize their key properties, which frequently involves the use of statistical graphics and other data visualization techniques. With the use of summary statistics and graphical representations, it aims to find patterns, spot anomalies, test hypotheses, and verify assumptions.

Variable	Description
price	Price in US dollars (326-18,823)
carat	Weight of the diamond (0.2-5.02)
cut	Quality of the cut (Fair, Good, Very good, Premium, Ideal)
color	Diamond color, from D (best) to J (worst)
clarity	A measurement of how clear the diamond is (I1(worst), SI2, SI1, VS2, VS1, WS2, WS1, IF (best))
table	Width of the top of diamond relative to widest point (43-95)
depth	Total depth percent = z/mean(x,y) = 2*z/(x+y)(43-75)
x	Length in mm (0-10.74)
y	Width in mm (0-58.9)
z	Depth in mm (0-31.8)

Table 4.1: The Study Variables

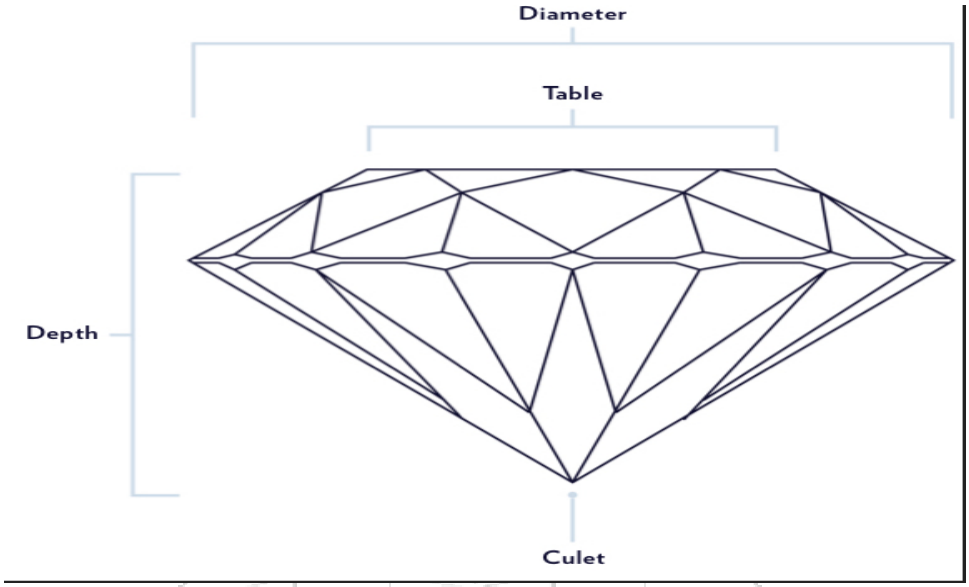

Figure 4.1: The Diamond's Key Measurements

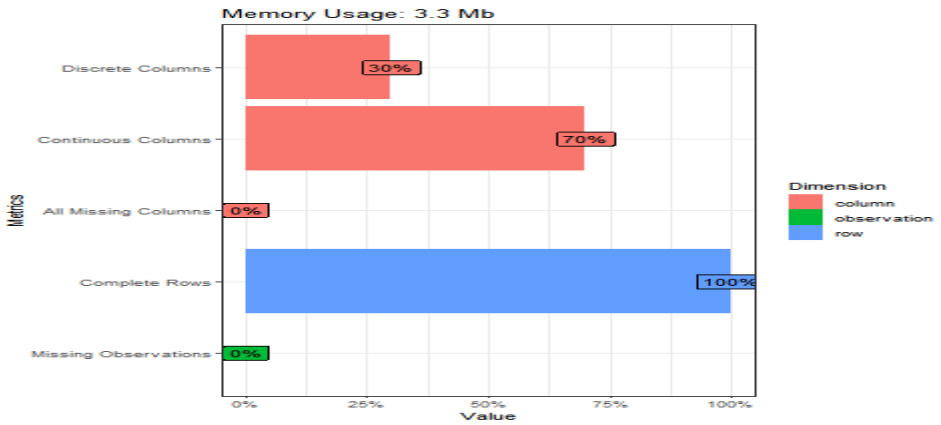

Figure 4.2: The Data Structure

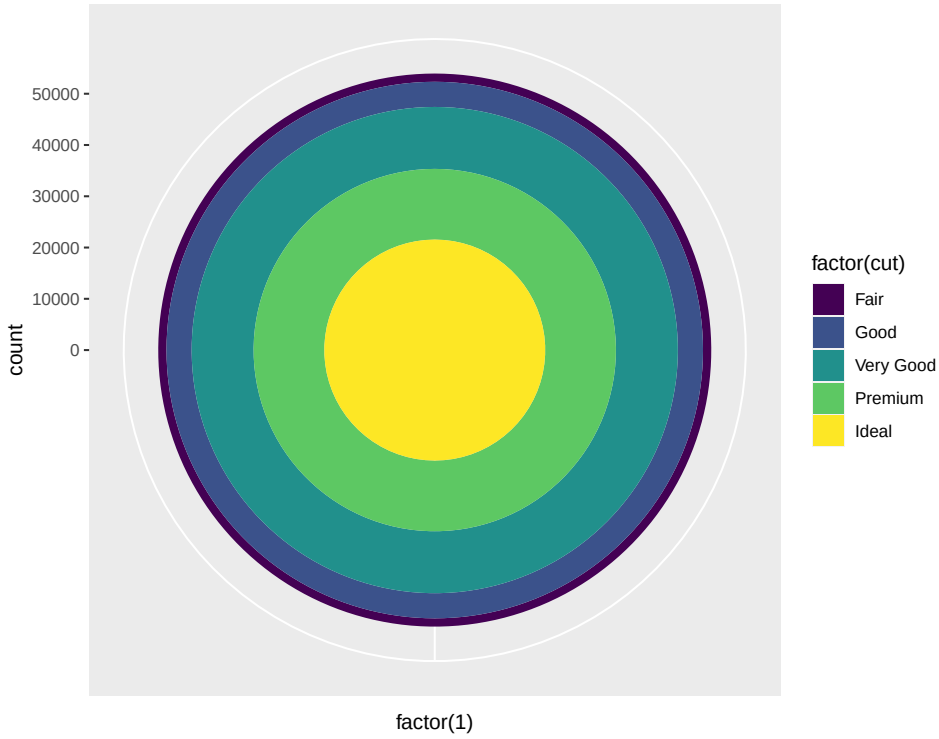

Figure 4.3: The Bulls-eye Chart

The *ideal* cut has the highest count while *fair* has the least as shown in Figure 4.3

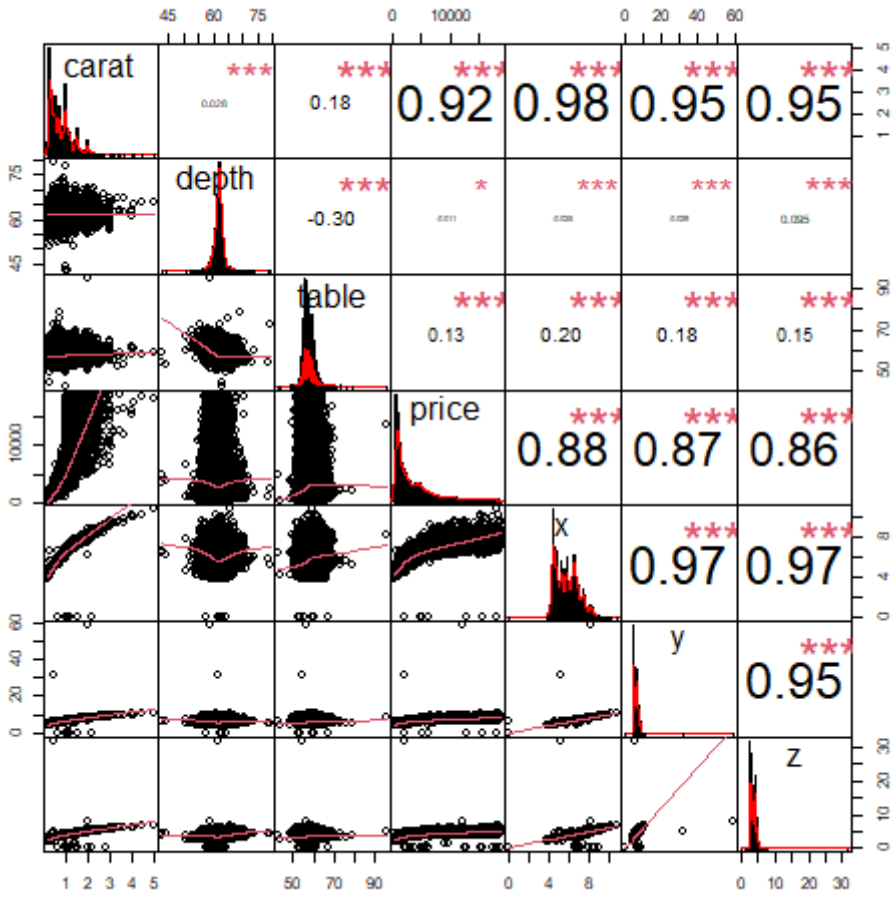

Figure 4.4: The Diamond Dataset Correlation Chart

Figure 4.4 indicates that the *carat* and *price* seem to be skewed to the right thus logarithmic transformation is necessary to achieve normality. Variables *depth* and *z* have normal distribution but more peaked than normal i.e. *leptokurtic*.

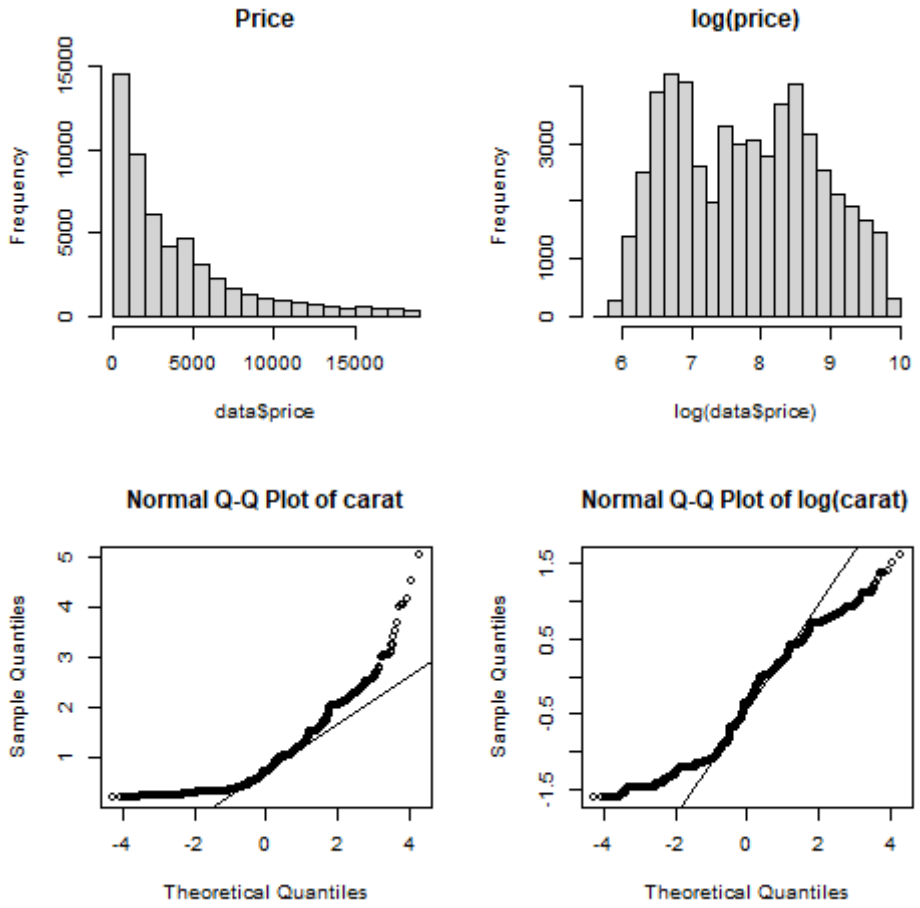

Figure 4.5: The Logarithmic Transformation of Price and Carat

Figure 4.5 indicates that the price and carat logarithmic transformation achieves normality.

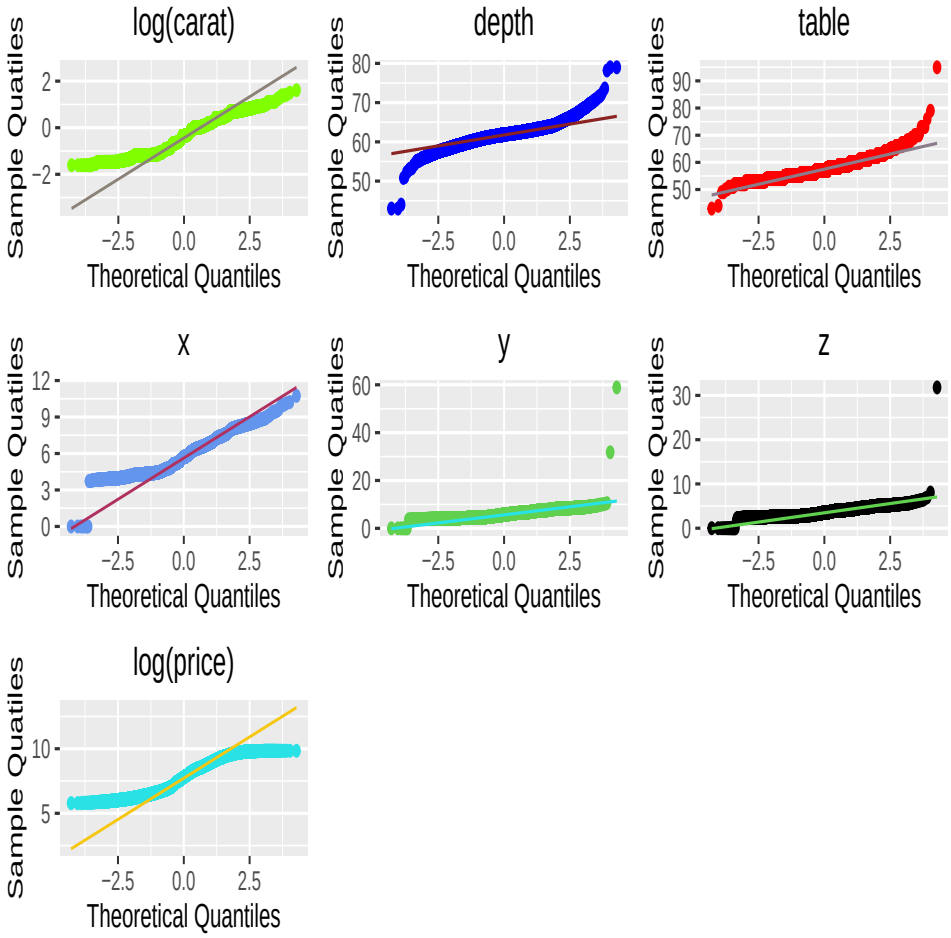

Figure 4.6: The Normality Test

Figure 4.6 shows that the normality assumption is generally satisfied save for slight dispersion for some variables caused by outliers which tend to draw the line of best fit to themselves.

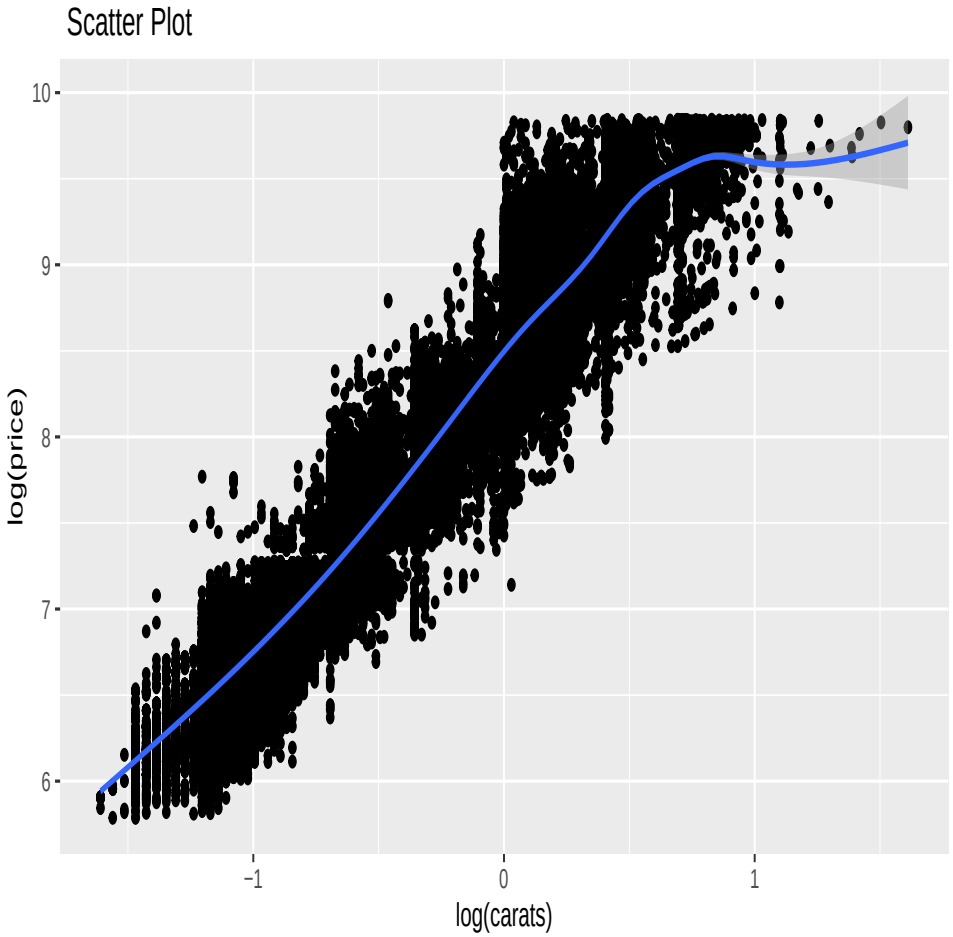

Figure 4.7: The Scatter Plot

Based on Figure 4.7, there is evidence of outliers in the data, as shown by a log of carat weight greater than one, where the confidence band begins to widen. It is therefore critical to remove them from the study in order to avoid producing spurious and nonsensical results and their interpretation. However, the removal of outliers is critical and should follow a judicious process, as they should only be removed if it can be proven that they are the result of error rather than natural causes. As a result, this study considers leaving the outliers alone.

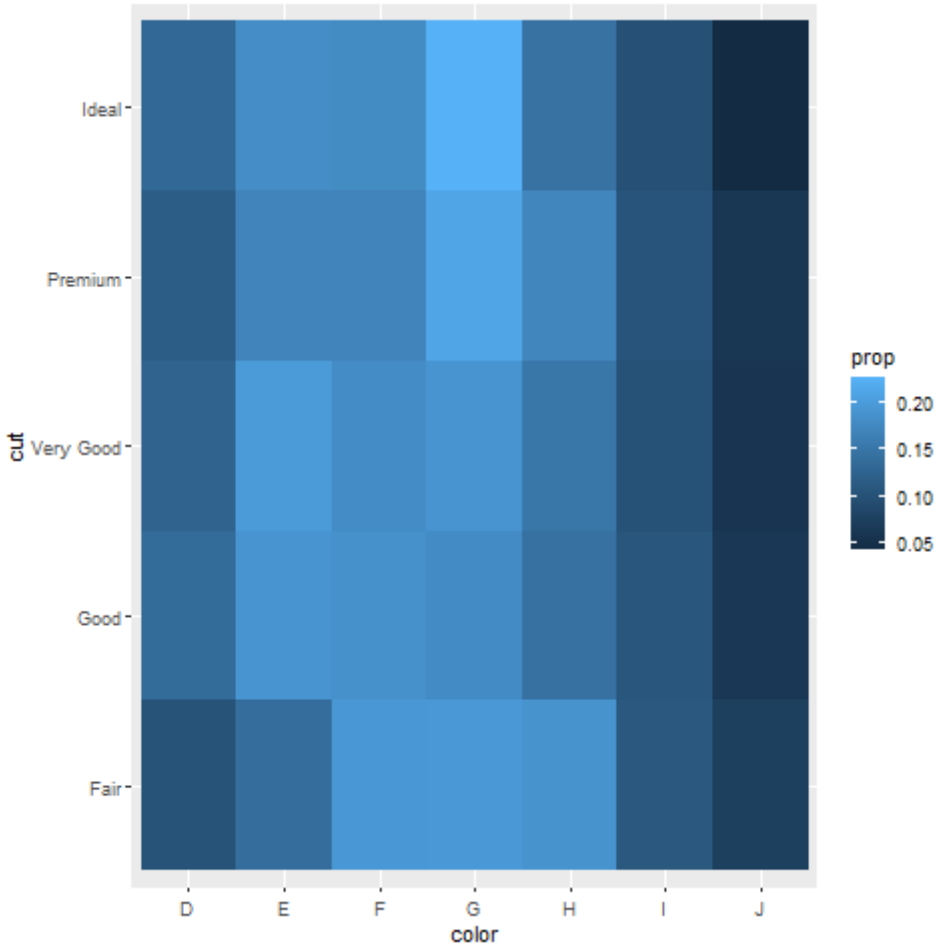

Figure 4.8: The Heatmap of cut and color

From Figure 4.8, we can conclude that:

- Most ideal and premium cuts are from colour G.
- Most very good and good cut diamonds are from colour E.
- Fair cut diamonds are usually from colour F, G, H.
- Overall, all cut group diamonds are rare in colour J.

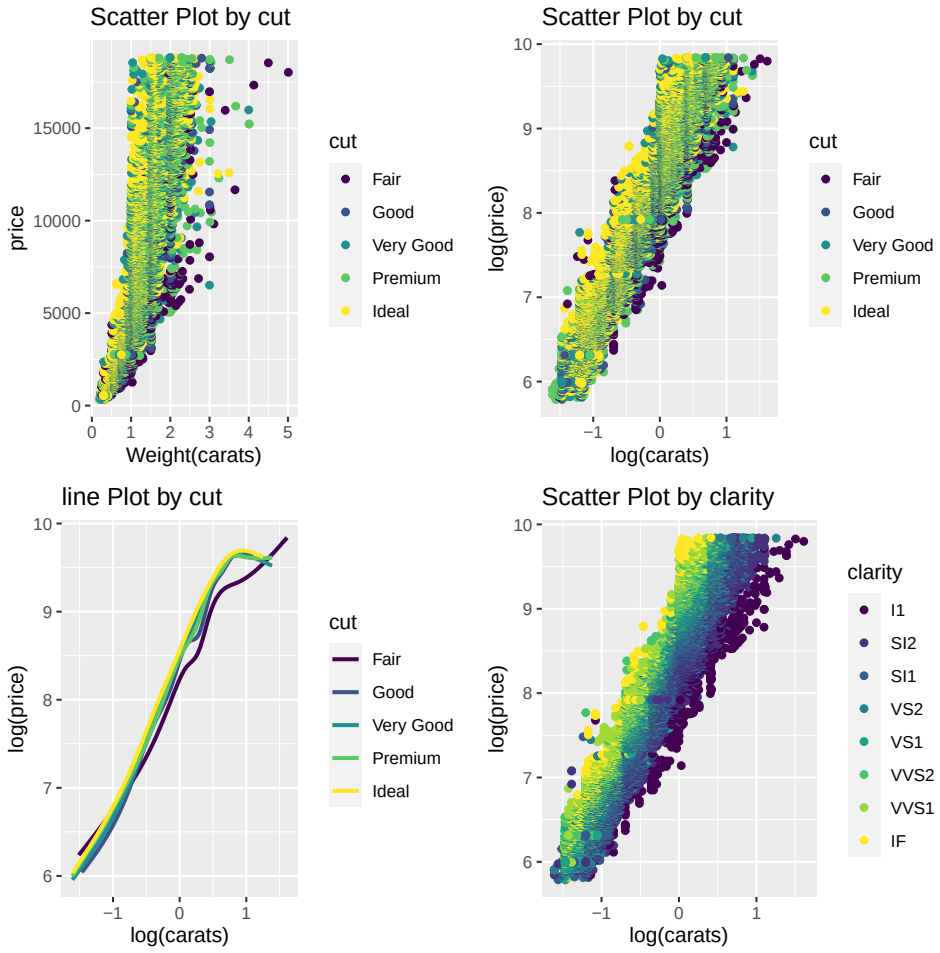

Figure 4.9: The 4Cs Visualizations

Figure 4.9 indicates that:

- The relationship between diamond price and weight(carats) proves to be relatively non-linear where heavy diamonds seem to exhibit higher price volatility, thus requiring logarithmic transformation.
- Diamond price seems to be an increasing function of the 4Cs.
- There is a positive correlation between the price and the carat for the different cut diamonds. This is confirmed on the correlation chart where price and carat exhibit correlation of 0.92.

Table 4.2: Regression Evaluation Metrics

Algorithms	R^2	RMSE	MAE
XGBoost	97.45	646.69	347.94
Rf	97.13	704.61	347.13
BRT	96.88	707.22	369.36
SVR	91.84	1164.68	662.91
kNN	85.35	1504.37	793.12

Table 4.3: Classification Evaluation Metrics

Algo's	Prec	Recall	F1	Kappa	Accuracy
XGBoost	71.89	77.26	71.63	63.06	74.28
BCT	73.37	71.44	72.04	63.12	74.09
Rf	74.51	71.49	72.10	61.00	72.61
SVM	75.19	55.62	55.45	52.03	66.95
kNN	53.30	41.80	43.95	36.96	55.8

Table 4.4: Algorithm's Lead Table

Algorithms	R^2	RMSE	MAE	Precision	Recall	F1	Kappa	Accuracy
SVM	4	4	4	4	3	3	4	4
BCART	3	3	3	2	2	2	1	2
XGBoost	1	1	2	4	1	3	2	1
Rf	2	2	1	2	1	1	3	3
kNN	5	5	5	5	5	5	5	5

Table 4.5: Overall Algorithms' Performance

Algorithms	Rating (%)
SVM	0.00
BCART	12.50
XGBoost	50.00
Rf	37.50
kNN	0.00

$$Overall\ Performance = \frac{X}{8} * 100$$

X=Total number of metrics.

Results in Table 4.5 indicate that the best ML model in regression and classification is *XGBoost* at 50%.

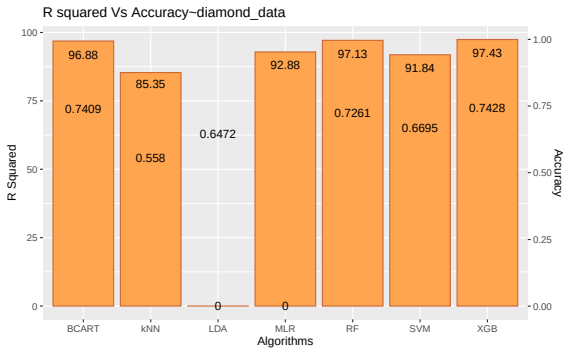

Figure 4.10: R squared Vs Accuracy

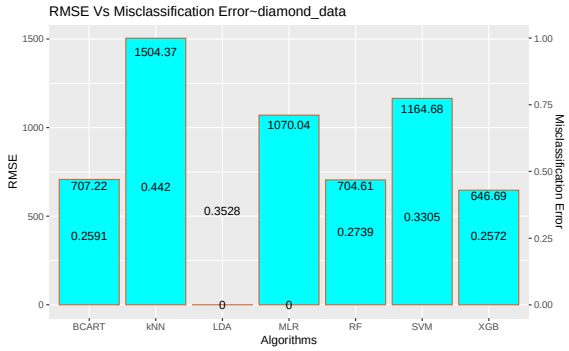

Figure 4.11: RMSE Vs Misclassification Error

With an accuracy of 74.28% and a R^2 value of 97.45%, Figure 4.10 shows that XGBoost is the overall best model for both classification and regression scenarios. BCART and Rf both perform well in classification (74.09% vs. 72.61%) and regression (96.88% vs. 97.13%), respectively. With RMSE of 1504.37 and 1164.68, respectively, kNN and SVM (Polynomial) have greater error values, making them poor regression predictors as indicated by Figure 4.11.

Chapter 5

Discussion, Conclusion and Recommendations

5.1 Introduction

The objective of this section is to interpret and discuss the importance of the study findings in connection to the research problem under investigation, as well as to explain any new knowledge or insights gained from the research. Diamonds dataset is used to train and validate all of the models discussed earlier in chapter three.

Here, the goal is to predict the price of diamond using key diamond features. The evaluation begins by dividing the dataset into two parts: the Train set (80%) and the Validation set (20%). Our model can make predictions on values it has never seen before thanks to the Validation set. All of the models under consideration were subjected to a k-fold crossvalidation, with k set to 5. The dataset was scaled and centered for feature comparability.

5.2 Discussion

5.2.1 Regression Evaluation Metrics

Price was regressed against nine other variables, including categorical ones, in regression (color, cut and clarity). One Hot Encoding (OHE) was used to convert these factor variables. OHE is a crucial step in the process of translating categorical data variables into machine

and deep learning algorithms, which improve model predictions and classification accuracy (Seger, 2018).

The XGBoost model outperformed all algorithms in terms of the regression metrics tested. For the R^2, RMSE, and MAE, the regression metrics scores were 97.45%, 646.69, and 347.94, respectively. These results were an improvement above the Alsuraihi et al. (2020) XGBoost Model, which achieved RMSE and MAE values of 1406 and 938, respectively.

In this study, the optimal model was achieved after tuning the critical architecture parameters as follows; (max.depth = 6,epochs = 46, eta = 0.3, gamma = 5, nfold = 5, booster = gbtree). It's worth noting that XGBoost seems to perform worse on tiny data sets, such as the simulated and iris datasets, which each comprised 3000 and 150 observations.

With a R^2 of 97.13% and cost function (RSME and MAE) values of 704.61 and 347.13, Rf was the second best performing model. Cross-validation (k-fold) was performed where k was held at 5. This result, although being based on a collection of 9 features, equaled Pandey et al. (2019), where features were reduced to only 5. In this case, the Rf R^2 score was found to be in close agreement with 97.93% by (Sharma et al., 2021).

Random Forest followed the same pattern as XGBoost in that performance was poorer on small datasets, such as iris and simulated datasets. On regression scenarios, kNN performed the poorest, with a R^2 of 85.35%, 1504.37 and 793.12 for RMSE and MAE, respectively while k (neighbors) was held at 2. All of the other datasets followed the same pattern for kNN regression.

5.2.2 Classification Evaluation Metrics

The response variable in classification was *cut*, which contains five classes (Fair, Good, Very Good, Premium, Ideal). XGBoost exhibits the highest Accuarcy and Recall in classification, at 74.28% and 77.26%, respectively. Here, the optimal model was achieved after tuning the critical architecture parameters as follows; (max.depth = 6, epochs = 40, eta=0.001, gamma=5, nfold=5, booster = gbtree).

The accuracy of the BCT model is 74.09% at (k-fold = 5, booster = xgbTree). While SVM (k-fold = 5, version = svmLinear) isn't the greatest at Accuracy (66.95%), it is the best at Precision (75.19%).

All of the classification evaluation measures show that kNN performs the worst at (k-fold = 5,neighbors = 17). However, with tiny datasets (irs and simulated datasets), it produces good results. kNN may not be the best model for evaluating huge datasets, as evidenced. At F1-scores, BCT and Rf perform best (72.04% and 72.10%, respectively). In terms of *kappa*, BCT, XGBoost, and Rf are at the top with 63.12%, 63.06%, and 61%, respectively.

5.2.3 Performance of Ensembles

We began by classifying modeling techniques into three categories: linear models, ensemble models, and others. XGBoost, which falls under the area of Ensembles and more precisely under the *Boosting Techniques*, was the best model for classification and regression as confirmed by key metrics i.e. a R^2 of 97.45% and an Accuracy of 74.28%. As a result, the findings show that Boosting outperforms Bootstrapped Aggregation (Bagging) in terms of prediction.

5.2.4 Algorithms' Overall Performance

On a small dataset (the iris dataset), kNN has the best overall classification and regression prediction performance at 28.57%. In larger datasets (simulation and diamond datasets), XGBoost takes the lead with 57.14% and 50%, respectively. This outstanding performance can be credited to XGBoost's cutting-edge architecture, which navigates large and complicated data structures and feature interactions.

5.3 Conclusion

The eXtreme Gradient Boosting (XGBoost) outperformed other algorithms in diamond classification based on cut, an alternative price prediction approach proposed in this paper. Furthermore, XGBoost generated the best results in diamond price prediction using regression, demonstrating that it is the best tool in diamond price prediction utilizing both methodologies.

5.4 Recommendations

5.4.1 Recommendations for Further Studies

Increase processing power to handle complex algorithms such as MLP by developing and deploying end-to-end GPU-accelerated data science workflows that allow for rapid exploration, iteration, and deployment of work. Using the RAPIDS-accelerated data science libraries, it would be feasible to execute data analysis at scale using a wide range of GPU-accelerated machine learning methods, such as XGBoost, cuGRAPH's single-source shortest path, and cuML's KNN, DBSCAN, and others. This study, for example, had to abandon the MLP approach, despite its promise of high output, due to a lengthy training time that was discovered to be computationally expensive.

5.4.2 Policy Recommendations

This study recommends creating an online interactive space, such as R-Shiny, where diamond attributes are fed and the model generates the most accurate cut category (key price determinant) and thus justifiable price estimate, to eliminate information asymmetry that propagates price obfuscation by various diamond retailers.

References

Shivam Agrawal. Analyze diamonds by their cut, color, clarity, price, and other attributes. *Diamond Competition*, 2017. https://www.kaggle.com/shivam2503/diamonds, Accessed on May 24, 2017.

Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar, and Hisham El-Shishiny. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5-6):594–621, 2010.

Waad Alsuraihi, Ekram Al-hazmi, Kholoud Bawazeer, and Hanan AlGhamdi. Machine learning algorithms for diamond price prediction. In *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, pages 150–154, 2020.

Anthony G Barnston. Correspondence among the correlation, rmse, and heidke forecast verification measures; refinement of the heidke score. *Weather and Forecasting*, 7(4): 699–709, 1992.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

Margarida GMS Cardoso and Luis Chambel. A valuation model for cut diamonds. *International Transactions in Operational Research*, 12(4):417–436, 2005.

Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 96–103, 2008.

Schmidt Chris. Analysis of LR, LDA, QDA, GAM models with K-CV. *RPubs*, 2021. https://rpubs.com/ChrisSchmidt/777478, Accessed on June 14, 2021.

Singfat Chu. Diamond ring pricing using linear regression. *Journal of Statistics Education*, 4(3), 1996.

Singfat Chu. Pricing the c's of diamond stones. *Journal of Statistics Education*, 9(2), 2001.

Donald Clark. How to choose a diamond. *Expert Buying Guide*, 2022. https://www.gemsociety.org/article/choosing-a-diamond/, Accessed on March 8, 2022.

Georgios N Dimitrakopoulos, Aristidis G Vrahatis, Vassilis Plagianakos, and Kyriakos Sgarbas. Pathway analysis using xgboost classification in biomedical data. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, pages 1–6, 2018.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

Abhijit Ghatak. *Deep learning with R*. Springer, 2019.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Christian Kampichler, Ralf Wieland, Sophie Calmé, Holger Weissenberger, and Stefan Arriaga-Weiss. Classification in conservation biology: a comparison of five machine-learning methods. *Ecological Informatics*, 5(6):441–450, 2010.

A Kassambara. Linear regression essentials in r. *Retrieved on http://www. sthda. com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r*, 2018.

Alboukadel Kassambara. *Machine Learning Essentials*, volume 1. Sthda, 2017.

Erik Lampa, Lars Lind, P Monica Lind, and Anna Bornefalk-Hermansson. The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. *Environmental health*, 13(1):1–17, 2014.

Stanislav Mamonov and Tamilla Triantoro. Subjectivity of diamond prices in online retail: insights from a data mining study. *Journal of theoretical and applied electronic commerce research*, 13(2):15–28, 2018.

M.Garside. Global demand value for polished diamonds by country 2019 . *Diamond Industry*, 2020. https://www.statista.com/statistics/894919/global-polished-diamond-demand-value-by-country/, Accessed on November 11, 2020.

M.Garside. Global diamond jewelry market value by country 2020. *Diamond Industry*, 2021a. https://www.statista.com/statistics/585103/diamond-jewelry-market-value-worldwide-by-region/, Accessed on November 15, 2021.

M.Garside. Global diamond jewelry market value 2010-2020. *Diamond Industry*, 2021b. https://www.statista.com/statistics/585267/diamond-jewelry-market-value-worldwide/, Accessed on November 15, 2021.

M.Garside. Diamond industry statistics and facts. *Diamond Industry*, 2022. https://www.statista.com/topics/1704/diamond-industry/#dossierContents__outerWrapper, Accessed on February 15, 2022.

Harshvadan Mihir, Manish I Patel, Soham Jani, and Ruchi Gajjar. Diamond price prediction using machine learning. In *2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4)*, pages 1–5. IEEE, 2021.

Mohammad-Reza Mohammadi, Fahime Hadavimoghaddam, Maryam Pourmahdi, Saeid Atashrouz, Muhammad Tajammal Munir, Abdolhossein Hemmati-Sarapardeh, Amir H Mosavi, and Ahmad Mohaddespour. Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and equations of state. *Scientific reports*, 11(1):1–20, 2021.

Mohanad Mohammed, Henry Mwambi, Innocent B Mboya, Murtada K Elbashir, and Bernard Omolo. A stacking ensemble deep learning approach to cancer type classification based on tcga data. *Scientific reports*, 11(1):1–22, 2021.

Gretchen G Moisen. Classification and regression trees. *In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582-588.*, pages 582–588, 2008.

Douglas C Montgomery and George C Runger. Multiple linear regression. *Applied Statistics and Probability for Engineers*, pages 410–467, 2010.

Blue Nile. choose your diamond. *Blue Nile Education*, 2022. https://www.bluenile.com/education/diamonds#:~:text=This%20video%20explains%20the%204Cs,characteristics%20of%20buying%20a%20diamond., Accessed on March 8, 2022.

FY Osisanwo, JET Akinsola, O Awodele, JO Hinmikaiye, O Olakanmi, and J Akinjobi. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3):128–138, 2017.

Avinash Chandra Pandey, Shubhangi Misra, and Mridul Saxena. Gold and diamond price prediction using enhanced ensemble learning. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–4. IEEE, 2019.

Thearasak Phaladisailoed and Thanisa Numnonda. Machine learning models comparison for bitcoin price prediction. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 506–511. IEEE, 2018.

Fernando Salazar, MA Toledo, E Oñate, and R Morán. An empirical comparison of machine learning techniques for dam behaviour modelling. *Structural Safety*, 56:9–17, 2015.

Frank Scott and Aaron Yelowitz. Pricing anomalies in the market for diamonds: evidence of conformist behavior. *Economic Inquiry*, 48(2):353–368, 2010.

Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

Garima Sharma, Vikas Tripathi, Manish Mahajan, and Awadhesh Kumar Srivastava. Comparative analysis of supervised models for diamond price prediction. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 1019–1022. IEEE, 2021.

Xiaowei Song, Arnold Mitnitski, Jafna Cox, and Kenneth Rockwood. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. In *MEDINFO 2004*, pages 736–740. IOS Press, 2004.

Thanasis Vafeiadis, Konstantinos I Diamantaras, George Sarigiannidis, and K Ch Chatzisavvas. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9, 2015.

Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

Jeffrey M Wooldridge. A note on computing r-squared and adjusted r-squared for trending and seasonal data. *Economics Letters*, 36(1):49–54, 1991.

Qingyao Wu, Yunming Ye, Haijun Zhang, Michael K Ng, and Shen-Shyang Ho. Forestexter: an efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems*, 67:105–116, 2014.

Zizhen Yao and Walter L Ruzzo. A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data. In *BMC bioinformatics*, volume 7, pages 1–11. BioMed Central, 2006.

Appendix A

A.1 Ethical Review Committee Report

Strathmore UNIVERSITY

30th May 2022

Mr Kigo Samuel,
samuel.kigo@strathmore.edu

Dear Mr Kigo,

RE: Assessing Predictive Performance of Supervised Machine Learning Algorithms

This is to inform you that SU-IERC has reviewed and **approved** your above **SU Masters'** research proposal. Your application reference number is **SU-IERC1352/22.** The approval period is **30th May 2022 to 29th May 2023.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-IERC

Cc: Prof Fred Were,
Chairperson; SU-IERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

A.2 Similarity Report

Curiginal

Document Information

Analyzed document	Assessing Predictive Performance of Supervised Machine Learning Algorithms_136851.pdf (D139242561)
Submitted	2022-06-03T20:10:00.0000000
Submitted by	
Submitter email	Samuel.Kigo@strathmore.edu
Similarity	1%
Analysis address	library.strath@analysis.urkund.com

Sources included in the report

SA	**MLReport_Group116.pdf** Document MLReport_Group116.pdf (D132354410)	2
SA	**5735819.pdf** Document 5735819.pdf (D29041156)	1
SA	**QRM_Thesis_Niels_Nijdam_2573944.pdf** Document QRM_Thesis_Niels_Nijdam_2573944.pdf (D109696012)	3