AFRICAN CENTER OF EXELLENCE

IN DATA SCIENCE (ACE-DS)

A MACHINE LEARNING APPROACH FOR CLAIMS RESERVING IN AUTO INSURANCE

A case of Reunion Insurance Company in Malawi

A Dissertation submitted in partial fulfilment of the requirements of the degree of Master of Science in Data Science in Actuarial Science

At the University of Rwanda

Supervisor: Dr. LEON FIDELE RUGANZU UWIMBABAZI

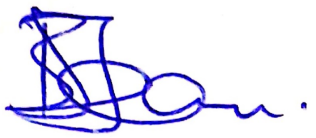
DONNEX JULIO BEYAMU

ACE-DS: 221000293

DECLARATION

I, **DONNEX JULIO BEYAMU (Reg. No. 221000293)**, do certify that this project is entirely original to me and has not been or will not be submitted to any other university for consideration for the same or another degree. The appropriate acknowledgements have been made whenever someone else's work or findings have been used.

Signed

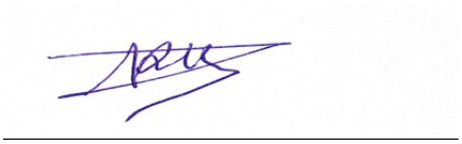


Date: 10th July, 2023

DONNEX JULIO BEYAMU (221000293)

APPROVAL SHEET

This dissertation, a machine learning approach for claims reserving in auto insurance: A case of Reunion Insurance in Malawi in partial fulfilment of the requirements of the degree of Master of Science in Data Science majoring in Actuarial Science is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 17% which is less than 20% accepted by ACE-DS.



Supervisor: **Dr. LEON FIDELE RUGANZU UWIMBABAZI (Ph.D.)**

Head of Training: **Dr. Ignace KABANO (Ph.D.)**

DEDICATION

I dedicate this thesis to my great and amazing parents, Mr. and Mrs. J. Beyamu, as well as to my four brothers, Amosi, Luciano, and Mannuel, and our family's lone sister, Triza. You are the best family in the entire world, so thank you for everything. God's blessings and my love to you!

ACKNOWLEDGEMENTS

First and foremost, I give God praise for the amazing benefits He has bestowed upon my life. I give Him praise for the strength, courage, and wisdom that saw me through times when things looked insurmountable. I praise God for His mercy and tenderness.

My sincere thanks and appreciation should be directed at my supervisor, Dr. Leon F. Ruganzu, for his unwavering encouragement, helpful criticism, concern, and understanding. His many suggestions have actually paid off; without him, my thesis would not have been as strong.

My parents, Mr. Julio and Mrs. Eveless Beyamu, deserve a special thank you for their unwavering love and support throughout my life, as well as for helping me socially, emotionally and spiritually. You have my undying affection and my prayers. Special thanks to Taona Kaima for her unwavering support when I was writing my thesis, as well as to my incredible sister Triza, my amazing brothers Amosi, Luciano, and Mannuel, and my cousin Lonjezo Mukongwa. You all made me smile, and I will always be grateful for you.

I would also like to thank Dr. D. Chapeyama, CEO of Reunion Insurance Company, and Mr. D. Malamba, my industrial Supervisor, for the support lendered to me during the journey of my academic internship at Reunion Insurance Company. To all Reunion staff team, I say thank you for the support and positive directions you provided me with during my internship period.

I want to express my gratitude to all of my fantastic friends and classmates, especially Euniter Chepchumba, Alexander Chidothi, Bernard Nangwiri, Amine Ahmed, Jean Claude Habimana, Javan Juma, and everyone else I got to know while I was studying at the University of Rwanda. Godspeed, everyone!

ABSTRACT

Accurate claims reserving is an important aspect for all insurers in the insurance business to meet their future legal liabilities. Despite many methods being proposed for prediction of claim reserves, it still remains a challenge in the industry and in particular, Malawian insurance industry is no exception. In recent years Machine Learning (ML) has caught many attentions across many industries, specifically ML has a vast range of applications in the insurance industry. This research aimed at investigating the use case of ML approaches namely; Random Forest (RF), Extreme Gradient Boosting (XGB) and Neural Networks (NN), on individual claim reserve predictions in auto insurance. Using scikit-learn and keras in Python, Random Forest, Extreme Gradient Boosting and Neural Network regressions were used on Malawian insurance claims data. Prior to model fitting, the data was exposed to various ways of preprocessing techniques such data cleaning, data transformation, feature selection and feature engineering where one-hot-encoding was used to transform all categorical variables into numerics. Then the data was splitted into two sets; one set for model training (80%) and the other for model validation (20%). In order to select the model that provides the most accurate claim reserve amounts in the auto insurance business, various evaluation criteria including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and accuracy score (R^2) were used. The results showed that RF achieved an accuracy score of 89%, MAE of 2.78, and RMSE of 3.91 while XGB and NN both achieved an accuracy score of 88%. Further, XGB achieved a MAE of 2.92 and RMSE of 4.11 while NN achieve a MAE of 2.91 and RMSE of 4.13. Thus, overall all the three models have produced desirable results as they have all achieved high accuracy scores. However, RF has slightly outperformed the other two methods and hence may be regarded as the best model out of the three. This implies that ML regression models, in particular RF, XGB and NN have the ability of producing the most accurate individual claim reserve predictions in auto insurance industry and they may therefore be considered in prediction of claims reserves.

***Keywords*:** Claims Reserving; Auto Insurance; Machine Learning; Random Forest; Extreme Gradient Boosting

TABLE OF CONTENTS

List of Tables

List of Figures

LIST OF ABBREVIATIONS

ML	Machine Learning
SVM	Support Vector Machine
NN	Neural Networks
DT	Decision Tree
RF	Random Forest
KNN	KNearest Neighbor
PCA	Principal Component Analysis
DL	Deep Learning
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
IBNR	Incurred but not reported
RBNS	Reported but not settled
XGBoost	Extreme Gradient Boosting
GLM	Generalized Linear Model
BFM	Bornhuetter-Ferguson Method
CLM	Chain-Ladder Method
ANN	Artificial Neural Networks

CHAPTER 1

INTRODUCTION

1.1 General Introduction

Claims reserving is the process of setting aside money for policyholders who have filed or are anticipated to submit genuine claims on their policies in the future. It is also referred to as loss reserving. Claims reserving estimation is one of the challenges facing the insurance business industry in the world Björkwall (2011). Due to the gap between the dates of the occurrence and the settlement, the insurer must set aside "reserves" for any claims that have not yet been resolved. The reserves needed at any one moment are the funds required to cover all costs associated with claims that have not been definitively resolved at that point Straub and Grubbs (1998). To accurately evaluate its financial condition for statutory and internal purposes, the insurer must be able to calculate this responsibility. According to Grace and Leverty (2012), most insurance companies either underestimate or overestimate the reserve which affects the profitability of the entire insurance business, despite it being an indispensable part of the functioning of every insurance business as it ensures the solvency of an insurance company.

Insurance companies safeguards policyholders against elements of likelihoods that may lead to financial losses to individuals and companies Gründl et al. (2017). They trade uncertainty for a certainty by pooling resources from a diverse group of individuals to help the few in case of eventualities. People pay premiums to insurance companies and in return insurance companies cover the costs of such people once they suffer unprecedented loss. A portion of these premiums are invested in other financial securities while some amount is reserved to meet future claim liabilities. In the monetary report of a general insurance company, reserves are an important element and actuaries have been in great use in setting these amounts for proper and accurate reserving in insurance companies.

Actuaries use experience in calculating reserve amounts and most of these amounts are not al-

ways accurate as actuaries use mostly traditional methods and other unexpected simple methods based on the chain ladder and link ratio approaches Wüthrich (2018). These methods mainly focus on estimating the aggregate claim reserves for incurred but not reported (IBNR) losses depending on the historical loss experiences with the assumption that past experiences will recur in the future. On the contrary, when it comes to the reported but not settled (RBNS) claims, the methods fall short. When a claim is filed the claims adjusters do not exactly know how much money may be required to settle the claim though at times they may use experience but the complexities of different claim cases makes it a challenging process.

Nowadays, with increase in computing power, companies have resorted for much more flexible methods for analyses in their institutions. Different advanced analytical methods are being used to make the most out of business data that comes in different formats such as structured and unstructured data. In particular, the insurance industry is one of the biggest business sectors that handles diverse amount of data in their daily operations ranging from underwriting, claims management and legal sections. With the coming in of machine learning models which offers a lot of flexibility based on the type of the problem and information available, most analytics have become easy and business institutions can make the most out of their available data for better data-driven decisions.

Generally, from the moment an insurance client files a claim, it takes about 30 days or more to settle the claim Mahlow and Wagner (2016). Claim adjusters spend considerable time investigating the legitimacy of the claim, however, they are required to reserve some amount that may be needed to settle the amount immediately. This amount is regardless of whether the claim will be legit or not. Such claims are referred to as reported but not settled (RBNS) claims. Although Wüthrich (2018) has illustrated on the use of some ML models such as trees in prediction of individual claims reserve, however he focused on the number of claims rather than the severity of the claims. Further to that, Baudry and Robert (2019) also used decision trees in predicting individual claims on simulated data which he compared the results to the tradition Chain ladder method results to compare the performances of the models. Thus, this study aimed at applying these flexible machine learning methods on real claims insurance data other than simulated.

1.2 Statement of the Problem

Solvency is a key factor for all insurance companies as every insurance company is expecting to fulfil its financial obligations to the policyholders insured under their policies. To accomplish this it is very pertinent that all insurance companies must have enough reserves to meet their future liabilities. While the aspect of having enough reserve capital is vital, the insurance company also needs to invest part of the premiums for maximising profits. It is therefore necessary to find a trade-off point so that the amounts should be a true reflection of the companies' future liabilities. Despite that literature has shown a massive number of contributions towards the prediction of claims reserves for individual claims, If any literature exists at all, it is sparse and pertains to the use of machine learning techniques for Malawian insurance industry. Therefore the study aims at addressing the gap by considering the use of these ML models in Malawian general insurance companies by studying a Malawian insurance company.

1.3 Research Objectives

1.3.1 Main objective

The main purpose for the study is to use ML techniques to predict individual claim reserves in auto insurance.

1.3.2 Specific Objectives

Specifically the study wants:

(a) To estimate claim reserves using ML regression algorithms.

(b) To analyse the collinearity between risk factors.

(c) To evaluate the association between the target variable and categorized independent variables.

(d) To compare the performances of different machine learning models on prediction of claims reserve.

1.3.3 Research Questions

Explicitly the research aims at addressing the following questions;

(i) What is the impact of risk factors such as accident severity, vehicle type, vehicle damage, e.t.c on claim amount?

(ii) Is there a relationship among the independent risk variables?

(iii) Which model has the best performance in prediction of claims reserves?

1.3.4 Research Hypothesis

H_0: There is a linkage between risk factors and individual claims reserve
H_1: There is no linkage between risk factors and individual claims reserve

1.4 Significance of the Study

As machine learning techniques have become an integral part of data analytics in different business industries, this study aims at investigating its use in the auto insurance industry in Malawi. The flexibility and accuracy of the models will allow insurance business to produce almost accurate estimates of their individual claim reserve amounts and this as a consequence will help to improve business operations for maximum returns.

Because they can measure claim reserves, insurance companies will be able to accurately assess the company's financial status, which is essential for both internal and legal purposes. Accurate individual claims reserves will help insurance companies to become solvent by fulfilling its legal liabilities in time. Further to that, the profitability of insurance business will grow by ensuring that the companies reserve accurate amounts and invest the other portion of premiums in other

investment financial securities for maximum returns.

Overall, this study will help insurance companies in Malawi and abroad in assessing the financial position of an auto insurance company, since a period's fluctuations in reserves are crucial for evaluating its progress. Additionally, by generalizing previous paid and reserved claim costs, insurance companies will be assisted in pricing their products in the sense of anticipating the future cost of claims on future risks.

CHAPTER 2

LITERATURE REVIEW

Introduction

This chapter aims at the analysis and evaluation of literature of the past researches about claims reserving in auto insurances and outlining theoretical and conceptual framework of the study. It explores further the traditional loss reserving techniques that have been used and how improvements have been made over time. The chapter further discusses different machine learning techniques that are useful in loss reserve predictions in auto insurance industry.

2.1 Review of Related Studies on Claims Reserving in Insurance

Over the years, various researches have been done to enlighten more about loss reserving in the insurance industry. Because of loss reservings' importance in the insurance industry, actuaries have devoted lots of their time in establishing ways how claims can be predicted before hand to ensure smooth operations of the business . Having accurate reserves at any particular moment in time is very vital for the wellness of an insurance company. Mack (1993) in his ground-breaking contribution, introduced Chain-ladder method (CLM) as a technique for calculating claims reserving in insurance. By extrapolating previous claim experience into the future, insurers utilize the chain ladder method to forecast the amount of reserves that it will need to be set up in order to cover anticipated future claims. Utilizing run-off triangles of paid losses and incurred losses, which represent the total of paid losses and case reserves, the chain ladder approach determines incurred but not reported (IBNR) loss estimates. Because of its simplicity, CLM is widely used in insurance companies. Despite the simplicity, CLM has a number of assumptions that needs to be satisfied for the predictions to be accurate. One of the speculations is that past loss patterns are projected to continue into the future. This implies that CLM does not generate an accurate

predictions without the necessary changes when an insurer's existing claims experience adjusts for some reason. Additionally, CLM only aims at estimating the claim reserve amount as an aggregated value and has less to tell about individual claim reserves.

Bornhuetter-Ferguson Method (BFM) is another loss reserving technique used in property and casualty insurance. It was introduced by Bornhuetter and Ferguson (1972). BFM is commonly seen as a combination of the chain-ladder and expected claims loss reserving methods as it employs an a priori estimated loss ratio and both reported or paid losses to arrive at an ultimate loss estimate Arico et al. (1972). Just like CLM method, BFM focuses to estimate incurred but not reported insurance claim amounts and estimates aggregate loss reserves. In contrary, Gabrielli (2021) argues that although working with aggregate data greatly simplifies the modelling process, but it also results in a significant loss of knowledge. The accuracy of claims reserving may be enhanced by using individual claims data. As a result, creating novel claims reserving methods that go beyond modelling aggregate data has gained popularity in actuarial science.

Later, with the high computing power, a number of researchers have explored new methods of calculating individual loss reserves. Taylor et al. (2008) carried out a research that used Generalised Linear Models (GLM) to calculate individual loss reserves where the emphasis was to look at the prediction errors associated with the loss reserves. Again, in 2019 Carrato and Visintin (2019) gave another loss reservation prediction method that blends conventional methods with ML techniques. Their goal was to strike a compromise between the traditional methods' ability to be interpreted and machine learning approaches' capacity for prediction. They recommended using machine learning to cluster the claims initially, such as using k-means or Gaussian mixtures, and then using conventional techniques to predict the reserves. Comparable claims would be placed in the same nodes on a single decision tree using the random forest method, which is similar to this.

Overall, there exist both deterministic and stochastic approaches for computing outstanding claims reserves, such as the chain ladder method and the Bornhuetter-Ferguson method which are based on aggregate claims data organized in claims development triangles Schmidt et al.

(2007). These methods have been used greatly in managing reserve risk for most insurance companies, however they are known that they can produce undesirable estimates if the underlying assumptions are not met. As suggested by Baudry and Robert (2019), some of the methods' shortcomings include the following: "a potential lack of robustness and the need for appropriate treatments of outliers; an over-parametrization risk caused by a large number of tail factors that must be estimated compared to the number of components of the run-off triangles; a risk of error propagation through the development factors associated to a possible huge estimation error for the latest development periods; and the impossibility of estimating the most recent development periods".

2.2 Machine Learning in Claims Reserving

This section introduces machine learning methods that are applicable in claims reserving prediction problems. Various types of ML are discussed and their typical examples of problems they can address are highlighted. However, as this section only tries to highlight ML that are useful in loss reserving problems, not all the discussed techniques were used to achieve the objectives of the study. The study only makes use of a few selected techniques to address the objectives of the study.

2.2.1 Definitions of Key Concepts of Machine Learning

Jordan and Mitchell defined Machine Learning as a discipline that focuses on how computer systems automatically keep improving through training experience when executing some type of tasks. For instance in predicting a loss reserve amount, the machine keeps learning from the past data and keeps on improving every time it is given a task to predict, in this case, the training experience is the policyholder's information such as driving experience, vehicle information, accident details and claim amount. There are three types of machine learning methods namely; supervised learning, unsupervised learning and reinforcement learning.

2.2.2 Supervised Learning

Supervised learning, often known as supervised machine learning, is a subset of machine learning and artificial intelligence. It differs from others in that it teaches computers to correctly categorize data or forecast events using labeled datasets. The algorithm is used to learn the function that changes input variables into output variables. Regression and classification are both supervised learning subclasses, according to Kennedy (2013). Regression works with continuous results, whereas classification deals with discrete outcomes like binary outcomes or those response variables that have more than two classes.In this study, supervised regression methods will be applied on the dataset. The supervised algorithms Random Forest (RF), Logistic Regression, Decision Tree, Support Vector Machine (SVM), Neural Networks, Extreme gradient Boosting and Linear Regression are the most well-known and often employed. Given that the dataset contains both the features and the response variable, this study will employ supervised techniques for regression.

2.2.3 Unsupervised Learning

Unsupervised learning is a sort of machine learning in which no outcomes are labeled and just the predictor input data is given. The main goal of this type of learning is to create models of the underlying data so that it may learn and comprehend more about it. Clustering and association are two subsets of unsupervised learning. Clustering is the process of identifying distinctive groupings in the data, such as classifying policyholders according to the level of their risk, whereas association is the process of identifying patterns or laws that broadly characterize the data Kennedy (2013), such as the tendency for people who buy insurance product B to also buy insurance product C. The widely used algorithm in this category is K-means.

2.2.4 Reinforcement Learning

The reinforcement system, referred to as an agent in this context, can observe the circumstance, choose and carry out actions, and receive positive or negative reinforcement as a result. The agent then discovers, on its own, the appropriate course of action, or policy, to maximize rewards over time. By acting appropriately, one can maximize rewards in a specific circumstance Patel

(2019).

2.3 Research Gap in Earlier Research

Due to setbacks of the traditional loss reserving methods such as the CLM and the BFM, a number of consideration of other different methods of calculating reserves have been studied England and Verrall (2002); Hindley (2017); Raeva and Pavlov (2017). Recently, various papers have been published for prediction of individual claims reserving which moves away from the aggregate claims reserving using better and advanced machine learning techniques which are flexible enough to cover diverse variations of risk factors. In addition, there has been an increasing growth of data referred to as big data and this requires new and improved methods of analysis for better data-driven insights.

On this note, different machine learning (ML) models have become handy because of their flexibility to capture various variations. These models are increasingly well-liked in data analytics and provide extremely accurate and flexible algorithms that can handle any type of data. Wüthrich Wüthrich (2018), in his contribution has suggested to an illustration of how the regression tree techniques can be used for individual claims reserving. However, his focus was only on the number of payments and not the severity losses paid. He further assumed that a homogeneous marked Poisson point process could adequately characterize the claim occurrences and reporting procedure, and as a result, these numbers of IBNR claims were forecasted by a CL approach that took use of the homogeneity assumption.

Baudry Baudry and Robert (2019) later considered other different machine learning methods where he focused on the individual claim severities. He used simulated data to assess the validity of the models and by contrasting the reserve estimate prediction values with the chain ladder for aggregate claims approach, the performance of his model was assessed. Again, recently Dong Qiu Qiu (2019) worked on the use of ML in loss reserving where the outcomes of applying ML techniques like Neural Networks (NN) and Random Forest (RF) were then contrasted with those of using traditional techniques on both simulated and actual data.

In general, this study intends to consider real general insurance claims data for reported but not settled claims and find out how advanced machine learning models such as Neural Networks, Random Forest and Extreme Gradient Boosting can be used in the calculation of individual loss reserves for RBNS claims from a Malawian Insurance company. As there are considerably many factors that contributes to the amount of claims required to settle a claim, the use of ML algorithms provides the best insights as to how much each of the factors contribute towards the claims amount. The flexibility ML models gives a boost in obtaining accurate claims estimates for auto insurance companies. This study was therefore aimed to address the gap by employing machine learning techniques which have the capabilities of producing most accurate results on real insurance data.

CHAPTER 3

MACHINE LEARNING MODELS AND METHODOLOGY

Introduction

This chapter aims at explaining machine learning models and the research methodology used through this project, the expected data to be used and its possible sources are also be outlined. The explanation of the data includes the type and expected target variable and explanatory variables. On top of that, the analysis methodology to meet the research objectives are outlined to stretch a clear picture.

3.1 Machine Learning Models

There are various ML models that can be used in insurance industry, both supervised and unsupervised learning models are more applicable. In a dataset, the relationship between the response feature and independent variable(s) is examined using the predictive modeling technique known as regression analysis. When the target and independent variables exhibit a linear or non-linear connection, and the target variable comprises continuous values, many forms of regression analysis techniques are employed. The major purposes of the regression technique are to assess predictor potency, forecast trend, time series, and in cases of cause-and-effect relationships. In particular, in claims reserving problem, various supervised learning models for regression are useful and this section discusses some of such ML regression models.

3.2 Machine Learning Regression Models

According to Arthur Samuel Samuel (1959), he defines Machine Learning models as mathematical models trained to make judgements or predictions without being explicitly programmed. Later Tom Mitchell Mitchell and Mitchell (1997) defined Machine Learning as the study of computer programs that get better on their own over time. There are different types of ML tasks

namely; Supervised ML, Unsupervised Ml and Reinforcement learning.

The target variable being numerical, only regression models were implemented. There many regression models which could have been used, however, only a few were selected. Particularly, the study used Random forest, Extreme Gradient Boosting and Neural Networks because of their ability to produce most desirable and robust results. These models also assume no distributions of the data and they can be used on non linear regressions. Hence being the ideal models for the type of data at hand.

3.2.1 Random Forest

Random Forest is an ensemble supervised machine learning technique that predicts the output using several decision trees Breiman (2001). Using this method, a decision tree is built using k randomly selected data points from the provided dataset. Then, several decision trees that forecast the value of any fresh data point are modelled.

Model Training

Precisely, given the dataset;

$$\mathbf{D} = \{(x_i, y_i), i = 1, 2, \cdots, n\}$$

where $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, we want to find a function $\hat{f}$ such that $\hat{\mathbf{Y}} = \hat{f}(X)$ for random variables of $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ representing the p-dimensional random vector and $\mathbf{Y}$ representing the response's real value. Then $\hat{f}$ is used in predicting the value of the response from the predictors: $y_0 = \hat{f}(\mathbf{x}_0)$ where $\mathbf{x}_0 = (x_{01}, x_{02}, \cdots, x_{0p})$.

The prediction function is defined by the loss function $L(Y, f(X))$ to minimize the predicted value of the loss. The expression is $\mathbb{E}_{XY}(L(Y, f(X)))$, where the subscripts stand for the expectation regarding the joint distribution of X and Y.

By penalizing $\hat{f}(X)$ values that are too distant from Y, $L(Y, f(X))$ is a straightforward way to

gauge how close $\hat{f}(X)$ is to Y. A popular option for L in regression is squared error loss:

$$L(Y, f(X)) = (Y - \hat{f}(X))^2$$

Below is a typical representation of a random forest algorithm is given in (3.1);

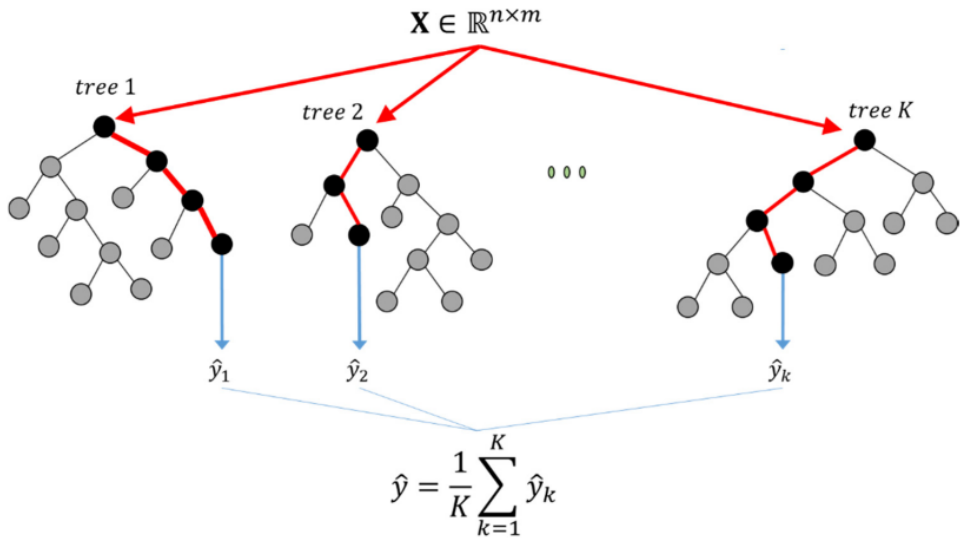


Figure 3.1: Schematic representation of a typical Random Forest; ***Source*** Aldrich (2020)

Random Forest employs the ensemble learning technique, which combines predictions from different machine learning algorithms (in this case different decision trees) to provide predictions that are more accurate than those from a single model (decision tree). The technique is referred to as the learning that uses the wisdom of the crowd principle rather than taking the decision of one model Sagi and Rokach (2018). A new data point's final result is calculated as the average of all the estimated values as seen in figure (3.1) above.

The most crucial Random Forest hyper-parameters that can be adjusted are: The number of decision trees in the forest (known as n-estimators in Scikit-learn) and the dividing parameters for each node (the MSE or MAE for regression). These parameters are either employed to accelerate the model or boost its predictive power Hastie et al. (2009). However, each hyper-parameter comes with its setback, for instance increasing the number of parameters increases the predictive power and model stability but it requires longer processing time.

3.2.2 Extreme Gradient Boosting

Another supervised machine learning technique that uses tree-based strategy in regression and classification predictive models is called Extreme Gradient Boosting (XGBoost). It is a strong and efficient application of the gradient boosted ensemble technique for forecasting in a variety of classification and regression issues. Decision tree models are used to build ensembles. In order to address the prediction mistakes made by earlier models, the trees are fed one at a time to the group and fitted.

XGBoost as first introduced by Tianqi Chen Chen and Guestrin (2016) uses advanced regularization (Lasso (L1) & Ridge (L2)), a more regularized variant of Gradient Boosting, to reduce over-fitting and increase the generalizability of models. Since XGBoost is an advancement of the gradient boosting approach, it performs better than gradient boosting. It parallelise training across clusters and is quite quick, this makes it one of the highly performing algorithm among Machine learning algorithms. Researchers love XGBoost for its execution speed, accuracy, efficiency, and usability Brownlee (2016).

Model Training

Given the dataset;

$$\mathbf{D} = \{(x_i, y_i), i = 1, 2, \cdots, n\}$$

where $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, and let the function $\hat{y} = \hat{f}(x)$ be the fitted weaker learner function of y such that $\hat{\mathbf{Y}} = \hat{f}(X)$ for random variables of $\mathbf{X} = (X_1, X_2, \cdots, X_p)$ representing the p-dimensional random vector and **Y** representing the response's real value.

XGBoost model as an extension of gradient boosting model, it has both the loss function as defined in the gradient boosting and regularization term which measures the complexity of the trees i.e. mathematically, the objective function is given as;

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.1}$$

where $l(y_i, \hat{y}_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i)^2$ (squared loss) is the training loss that measures how well the model fit on the training data and $\sum_{k=1}^{K} \Omega(f_k)$ is the regularization term. The regularization term $\Omega(f_k)$ has two parameters, the number of trees and the $L2$ norm of leaf scores expressed as;

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{3.2}$$

Where γ and λ are hyper-parameters. In the objective equation (3.1), the term $\hat{y}_i$ is given as the combination of the previous model and the new model i.e. $\hat{y}_i^{\{t\}} = \sum_{t=1}^{T} f_t(x_i) = \hat{y}_i^{t-1} + f_t(x_i)$. Then the objective function can be re-expressed as follows;

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.3}$$

Using the taylor series expansion on this objective function (3.3), Tiangi Chen and Guestrin (2016) expresses an approximation of the loss function as;

$$\text{Obj}^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{k=1}^{K} \Omega(f_k) \tag{3.4}$$

Where $g_i = \partial_{\hat{y}^{(t-1)}}(y^{(t-1)} - y_i)^2$ is the first order differentiation and $h_i = \partial^2_{\hat{y}^{(t-1)}}(y_i - \hat{y}^{(t-1)})^2$ is the second order differentiation of the taylor series expansion.

Now putting together the training loss (3.4) and the regularization term (3.2), the objective function is simplified to;

$$\text{Obj}^{\{t\}} = \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2}(H_j + \lambda)\gamma_j^2 \right] + \gamma T \tag{3.5}$$

Where $G_i = \sum_{i \in I_j} g_i$ and $H_i = \sum_{i \in I_j} h_i$ Chen and Guestrin (2016). For each quadratic function $G_j \omega_j + \frac{1}{2}(H_j + \lambda)\gamma_j^2$, gives an optimal weight of $\omega_j^* = -\dfrac{G_j}{H_j + \lambda}$, and substituting this into the objective function, we get a minimum objective function as;

$$minObj = -\frac{1}{2} \sum_{j=1}^{L} \frac{G_j^2}{H_j + \lambda} + \gamma L \tag{3.6}$$

Precisely, for t in $(1, 2, \cdots, T)$ sequentially add $f_t(x_i)$, learn the structure of $f_t(x_i)$ and get the minimal objective function which measures how good the tree structure is, given as in equation (3.6) above Chen and Guestrin (2016). Then the training stops either when the fixed T is reached, which represents the number of trees, or the objective reaches an acceptable error level. Then the final prediction model which is an additive combination of all the previous models built in the for loop, and it is given as;

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i) \tag{3.7}$$

The convex loss function (based on the difference between the anticipated and target outputs) and the model complexity penalty term are combined to form a regularized (L1 and L2) objective function that XGBoost minimizes.

3.2.3 Neural Network

Neural Network (NN) is another supervised machine learning algorithm that uses the idea of neurons, analogous to how the brain works. According to Kundar Kumar and Rao (2018), Neural networks also referred to as Artificial Neural Networks are computing systems under deep learning as a subset of machine learning. In the subject of machine learning, neural networks' behaviour, which mimics that of the human brain, enables computer systems to identify patterns and resolve frequent issues.

A Neural network is comprised of three sets of layers; the input, hidden and output layer, where each layer has nodes which connects across. The starting layer is known as the input layer which feeds the system's initial data to later layers of artificial neurons for processing. The algorithm's input and output are separated by a hidden layer, in which the function gives the inputs weights and sends them through an activation function as the output. Desirable predictions are made in the output layer, the last layer of the neural network. The desired result is produced by a neural network's single output layer. Prior to determining the final output, it applies its own set of weights and biases. Resultant outcomes depend on the activation function specified.

Model Training

Given the dataset;

$$\mathbf{D} = \{(x_i, y_i), i = 1, 2, \cdots, n\}$$

We need to find an approximation function $\hat{f}_{\mathbf{W}}(X)$ such that $\hat{f}_{\mathbf{W}}(X) = Y$, where $W = [W_0, W_1, \cdots, W_k]$ is a vector of weights, **X** is a a random vector of predictor variables and Y is the target value. Concisely, we seek a vector of parameters: $W = [W_0, W_1, \cdots, W_k]$ such that we have a linear relation between the response variable Y and predictor attributes X. The relationship mat be represented as;

$$\hat{\mathbf{f}}_{\mathbf{W}}(X) = W_0X_0 + W_1X_1 + \cdots + W_kX_k = \sum_{i=0}^{i=k} \mathbf{W}_i\mathbf{X}_i \tag{3.8}$$

The neural network model selects the weights **W** in a such a way that the error term is reduced. This is achieved by use of two principles known as the forward propagation and the backward propagation. In a forward propagation the weights which are associated to each variable feature in the data i.e. for features X_1 and X_2 weights W_1 and W_2 are randomly initialized and each weight is multiplied by the corresponding feature variable and these are served as input to a neuron. The neuron performs both summation, where all the weights are multiplied by their corresponding features and bias is summed up, and an activation function operates on the summed function to produce an output in the output layer. The randomly initialized weights then are readjusted in the second principle called backward propagation. This process of forward and backward propagation goes on and on until weights that give the most smallest error value are obtained. Below is graphical representation of a typical neural network model;

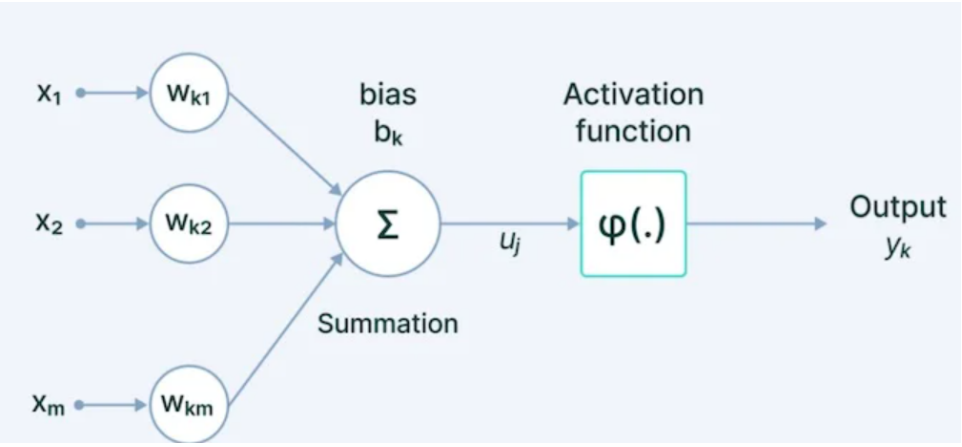


Figure 3.2: Schematic representation of a typical Neural Network model (*source; internet*)

Although this model is hard to interpret and implement because of its complexity and try and error method of choosing number of nodes and hidden layers, it was fitted because it learns well on large number of features with non-linear relationship between the independent features and dependent one.

3.3 Methodology

3.3.1 Data and Variable Description

As the study aimed at investigating the applicability of machine learning regression models on individual claim reserve predictions on Malawian data, secondary data was obtained from one of the general insurance companies in Malawi. The data which spanned from 2019 to 2022 consisted of 2048 rows and 17 columns, and out of 2048 cases only 970 were closed cases as they were all fully settled, while the remaining 1078 were partially settled and they were still open. For the purposes of the objectives of the study, only the closed claims were valid and as such the study used the 970 cases only. Further, insurance data is very confidential and sensitive as such the variables containing personal details were represented with temporary identification data values. Below are the data variables and their descriptions;

Table 3.1: Dataset Variables and their explanations

Variable	**Description**	**Data type**
Claim No.	Claim number assigned to the claimant and it serves as a claim ID	Integer
Policy No.	Insurance policy number for the client which is assigned to the client at the time the policy was bought	Integer
Policyholder	Alias names representing the policyholder	String
Registration number	Registration number of the claimant	String
Accident date	Date of the accident representing when the accident happened	Date
Description	Description of what exactly happened during the accident and including the severities of the damages	String
Vehicle Make	Make of the Vehicle involved in the accident	String
Model	Vehicle model for the claimant	String
Body type	Type of the vehicle body	String
Manufactured date	Year indicating when the vehicle was made to show how old the vehicle might be	Integer
Seat Capacity.	The number of passengers a vehicle can accommodate, including the driver as determined by the manufacturer.	Integer
Class	Alias name representing the Name of the policyholder	String
Motive power	Fuel type used by the vehicle i.e. whether the vehicle uses petrol or diesel	String
Colour	Vehicle color of the claimant	String
Chassis number	Vehicle chassis number	String
Engine Number	Vehicle engine number	String
Severity	Level of damage to the vehicle or individuals involved in the accident	Object
Vehicle damaged	Indicator whether the vehicle was damaged or not	Object
Body injuries	Whether there were any bodily injuries registerd	Object
Affected particulars	Explaining which parties that were affected in the accident	Object
Total claim payment	Total claim amount paid to settle the claim	Float

The target variable is the total Claim amount whereas the rest of the variables are the explanatory. The research is a quantitative design as the target variable is numerical, while the independent variables are both categorical and numerical. All the categorical features were transformed into numericals for better modelling. The results are based on evidence from the correspondent fitted models and other machine learning analytical techniques. Unlike qualitative researches, this project is not based on insights, experience and opinions. Data cleaning and Analyses were run in Python which was used because of its flexibility and simplicity in implementation of diverse machine learning models.

3.3.2 Data Preprocessing

Most of the dataset is not always clean therefore it is very important to make sure that prior to model fitting the data has to be checked for the inconsistencies such as missing values, null values, outliers and among others. The presence of these inconsistencies affect the performances of the fitted models making the models less meaningful and worthless. Thus, before fitting the models, the data was subjected to a series of preprocessing methods to make sure that it was in a format easy for model fitting to optimize efficiency of the models.

Machine learning models are sensitive to missing values such that having missing values may negatively affect the learning process of these models and consequently, this affects the overall performances of the models. Various ways of dealing with missing values which varies from one problem to another have been widely used. Complete deletion of the rows with missing values is one of the way of dealing with missing values problem. This method however depends on the size of the dataset in usage and it is not a best method when the data is already small as deleting missing rows may further reduce the size of training data available. For the same reason, this option was not used as part of the solution in the study as the available data was comparatively small hence reducing it further would have affected efficacy of the models training. The second method for handling the missing values is to impute values and there are different imputation methods some of which are, replacing all the missing values with zeros, means , mode or median of each respective variable feature. Since no data is lost, this method of handling missing data reserves the size of the dataset. For this reason, this approach was used to handle all the missing

data points and in particular all the missing values were replaced with the means of each variable feature.

Additionally, the data was checked for presence of outliers. Outliers which refers to extreme data points in a dataset do have a very negative effect when it comes to model fitting. They affects generalization of models leading to poor model performance. To make sure that all the outliers were identified and removed before model fitting, a percentile method was used based on the target feature, Total claim payment. All the data points that fell below thresholds of $Q1 - 1.5IQR$ and above $Q3 - 1.5IQR$, where $Q1$ represents the first quarter percentile, $Q3$ represents the third quarter percentile and IQR is the interquartile range given as the difference between $Q3$ and $Q1$ i.e. $Q3 - Q1$, were all considered as outliers and they were removed. Out of 970 cases, 31 were identified as outliers representing 3.2% of the entire valid dataset. This reduced the dataset to 939 valid cases for model training and testing.

3.3.3 Feature Selection and Feature Engineering

Feature selection

Feature selection as defined by Ha and Nguyen (2016) is a technique for selecting a subset of relevant features that can reduce dimensionality, improve model accuracy, and shorten runtime. Further to this, models needs to follow the principle of parsimony, which states that "a simpler model with fewer parameters is favoured over more complex models with more parameters, provided the models fit the data similarly well" Sober (1981). Despite many features being gathered in the dataset, not all the features are necessary in predicting the target variable, total claim amount of an individual. For instance, features such as those specific to the personal details of the client such as claim number, policyholder name, policy number, registration number, and others that had no relationship to the target feature such as chassis number, engine number, vehicle colour and description, are not useful in predicting the amount that a client will claim. Because of this, those features were dropped from the set of predictor variables.

Further, a plot of correlation between the target variable and the predictor variables plotted in Python and those features with significant correlation coefficients with the target variables

were considered important than those that had very small or close to zero coefficient values. The results of the correlation plot showed that those predictor features specific to the accident details of the claimant such as affected individuals, accident severity, personal injuries and vehicle damages, were highly correlated with the target variable of interest. Other features that showed significant correlation with the target feature were manufactured date of the vehicle and seat capacity. On the contrary, features such as model, make and body type showed coefficients values of close to zero. In reality one would expect that for different vehicle models the claim amounts also need to differ, under ceteris paribus assumption. Consequently, all features that showed significant correlations with the total claim amount were considered and the rest were dropped.

Feature Engineering

Machine learning models requires all features to be numerical in nature, this ensures that the data is in compatible with the statistical analysis under which machine learning models are based. Since most of the variable features such as severity, vehicle damaged, affected particulars and body injuries registered were all categorically recorded, there was a need to transform them into numerics to allow the fitted models to understand them perfectly. To achieve this, it required the use of a technique called feature engineering. Feature engineering in ML is a technique where data values are manipulated and transformed into features that can be used in supervised learning. This technique helps to improve the ML model training and consequently it leads to a greater accuracy and better model performance. One hot encoding was employed to convert all the categorical variables into numericals of dummy zeros and ones. On the other hand, numerical features such as seat capacity, vehicle manufactured date and the target variable, total payment were transformed into a range of values between 0 and 1 using Min-Max scaler package in python. This allowed to put the data points into a fixed range of values which makes model training easy and fast.

3.4 Analysis

3.4.1 Association and Correlation

Before carrying out model fitting, correlation analysis between the features and association between the features and the target variable was done.

Pearson Correlation

The Pearson correlation coefficient is denoted by ρ is given by the following formula:

$$\rho = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} \tag{3.9}$$

Where ρ is the coefficient of correlation, x and y are features being analysed and $\overline{x}$ and $\overline{y}$ are their means, respectively.

Chi-Square test for association

This one the tests the relationship between the categorical response feature and the other variables. To effectively implement it, the features would have to be categorized and encoded. Chi-square test of association literally works with categorical variables. The test statistic is calculated by the following:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{3.10}$$

Where O represent the observed frequencies and E represent the expected frequencies.

3.4.2 Hyper-parameter tuning and Models Optimization

Machine learning offers diverse ways of achieving the optimal model with the highest accuracy. Hence, after fitting Random Forest and Extreme Gradient Boosting models, Grid-Search and Randomized-Search-CV were used to find the optimal parameters for the models to achieve the best optimal prediction results. The optimal parameters obtained from the Grid-Search and Randomized-Search techniques achieved almost the same prediction results with negligible differences.

3.5 Evaluation Metrics

Different measures are used to assess how successfully machine learning regression models accomplish the tasks for which they were designed. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are typical metrics to be employed. These measures can be used to gauge how closely the projected values match the actual values.

3.5.1 Mean Squared Error

The simplest metric is Mean Squared Error (MSE) and yet mostly utilised in machine learning for evaluation of predictive regression models. Additionally, it serves as a substantial loss function for models that are fitted or improved using the least squares formulation of a regression issue Brownlee (2021). The average of the squared deviations between the estimated and anticipated target values is used to calculate MSE;

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3.11}$$

For N representing the number of data points, y_i is the expected real target value and $\hat{y}_i$ is the its predicted value, and MSE is a non negative number. For a best fitted regression model, MSE is expected to be a small value while for a poorly fitted regression model the MSE is very big.

The MSE gives these errors more weight because of the squaring component of the function, therefore it is useful for ensuring that the trained model does not have any outlier predictions with big errors. The squaring portion of the function, on the other hand, raises the error if the model makes a single, extremely poor forecast. However, in many real-world situations, we do not pay much attention to these outliers and instead strive for a more well-rounded model that works well enough for the majority of the data. Seif (2019).

In general, a better MSE is specific to dataset. The baseline MSE for the dataset will be obtained from the training MSE value. A model that gives a testing MSE that is small and close to the one

obtained from the training dataset will be regarded as the best.

3.5.2 Root Mean Squared Error

As an extension to the MSE, another metric for evaluating regression models is the Root Mean Squared Error (RMSE) .i.e. RMSE just as the name suggests it is a square root of MSE. Mathematically, RMSE is represented as;

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.12}$$

Or equivalently RMSE can be re-expressed in terms of MSE as;

$$RMSE = \sqrt{MSE} \tag{3.13}$$

Where N indicates the data points, y_i is the expected real target value while $\hat{y}_i$ is its estimated value. The square root in RMSE eliminates the error that was aggravated due to the squaring of the squared differences in MSE. Just like MSE, RMSE is also a non negative numeric value.

An ideal RMSE value of 0.00 indicates that all forecasts completely matched the predicted values. On the contrary, this is a relatively uncommon occurrence, and if it does, it shows that the problem with predictive modelling is straightforward..

In general, an excellent RMSE is specific to the dataset. The baseline RMSE for the dataset will be obtained from the training dataset values. A model that gives a testing RMSE close to the one obtained from the training dataset will be considered the best.

3.5.3 Mean Absolute Error

Since the units of the error score and the expected target value coincide, Mean Absolute Error (MAE) is a well-liked metric. Contrary to the RMSE, the changes in MAE are linear and logical. MSE and RMSE inflate or exaggerate the mean error score by penalizing greater errors more severely than smaller errors. This is because of the erroneous value's square. The scores in the

MAE rise linearly with the magnitude of the error rather than being weighted differently for various sorts of faults.

The absolute error numbers are averaged to determine the MAE score and it is always a positive number just like MSE and RMSE. Mathematically, MAE is given as;

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{3.14}$$

Thus, the distance between the model's estimates and the expected values can be used to determine MAE by applying the absolute value to the difference and then averaging it across the whole dataset.

In contrast to the MSE calculation, the MAE has the advantage that since the absolute value is taken, all errors will be given the same linear weighting. As a result, unlike the MSE, we won't place undue emphasis on our outliers, and the error function offers a general and consistent indicator of how well the strategy is doing Seif (2019).

On the other hand, the MAE will be less useful if we are concerned about our model's outlier predictions. Large errors from the extremes end up having the same weight as smaller errors. This could lead to our model performing well most of the time but occasionally producing some extremely inaccurate predictions.

3.6 Data Splitting to Train and Test Sets

After applying data preprocessing techniques such as data Cleaning/Cleansing, data Integration, data dransformation, and data Reduction, we divided the data into training and testing sets so that we could evaluate how well the models we had developed on the basis of the data performed. The models are fitted using the training data. The test data is the data that is utilized for an unbiased evaluation of the model fit on the training sample because the model observes and learns from this specific data and can be verified on it. The train-test split assesses how well the model generalizes and how well it responds to fresh input. When training the

data, it's crucial to keep an eye out for a number of factors, including optimizing the model for performance and checking that the models produced have the best generalizability to previously unexplored data.

We selected to train the model using 80% of the data for training and 20% that was randomly partitoned for an unbiased evaluation. More data points for training the models aid in the learning of the model. The formation and construction of the regressor model takes place during the training step, which is a crucial stage. This task was achieved by using a *traintestsplit* package in python and below is the table showing how the data was split.

Table 3.2: Train-Test Dataset Partitioning

Data	**Rows**	**Partition percentage**
Training Set	762	80%
Testing Set	191	20%

Data splitting in machine learning is very important stage before fitting ML models. Under this, the data was randomly divided into two different sets, one of which was used for training the model while the other set is for model validation which demonstrates how well the model works with fresh, unseen data. Hence using the train test split package in python, the data was split into training and testing set, where the training data was used to fit the models and later the testing data was used to assess the performances of the fitted models. This helped to evaluate the overall performances of the models on new unseen data. 80% (762 cases) of the data was used for training while the remaining 20% (191 cases) was used for testing and comparisons were made for different evaluation metrics against each fitted model.

CHAPTER 4

DATA ANALYSIS, RESULTS AND DISCUSSIONS

This chapter explains in details the main findings of the analysis. Firstly, exploratory data analyses will be presented together with all data preprocessing techniques that were applied on the data to make it good for model fitting. Thereafter, feature engineering will be done to ensure the variables are in a form that my be used in fitting models. Lastly, different ML models will be fit to the data to predict individual claim amounts.

4.1 Exploratory Data Analysis

Before the fitting regression models to the data, a number of exploratory data analysis were done to check the statistics of the data and how both the predictor variables (4.1) and the target variable are distributed. This inspection helped to see what if the proposed models were the right choices in addressing the objectives of the study. Further, the target variable was also explored independently and it was noted that the target variable was not normally distributed as it had a long positive tail.

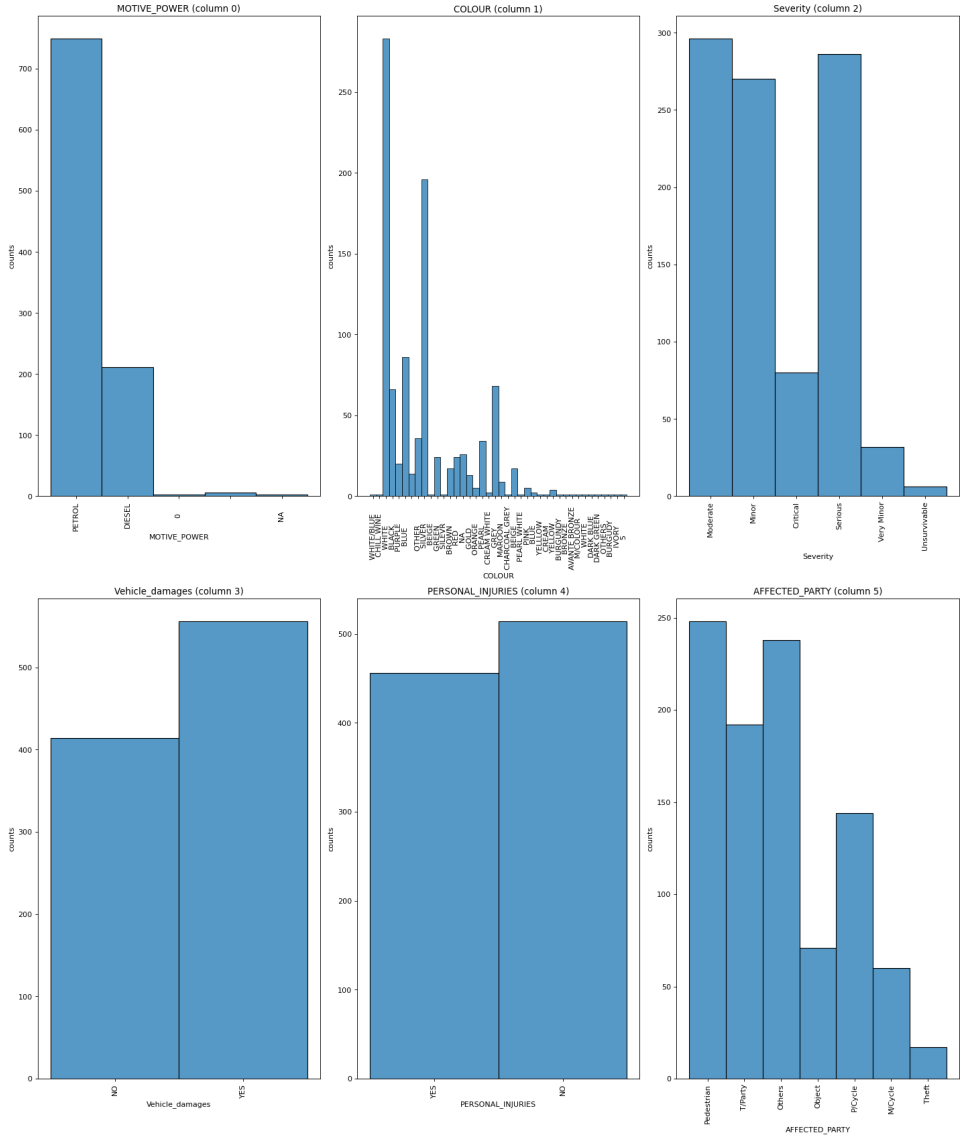


Figure 4.1: Distribution of independent variables

As seen the variables are categorical in nature hence the need to do feature engineering to transform them into numerical variables. The figure above (4.1) shows how different categories of of the independent variables are distributed.

Further to that another plot was done to check for the distributions of the numerical variables such as the manufactured date, seating capacity and the target variable, total payment.

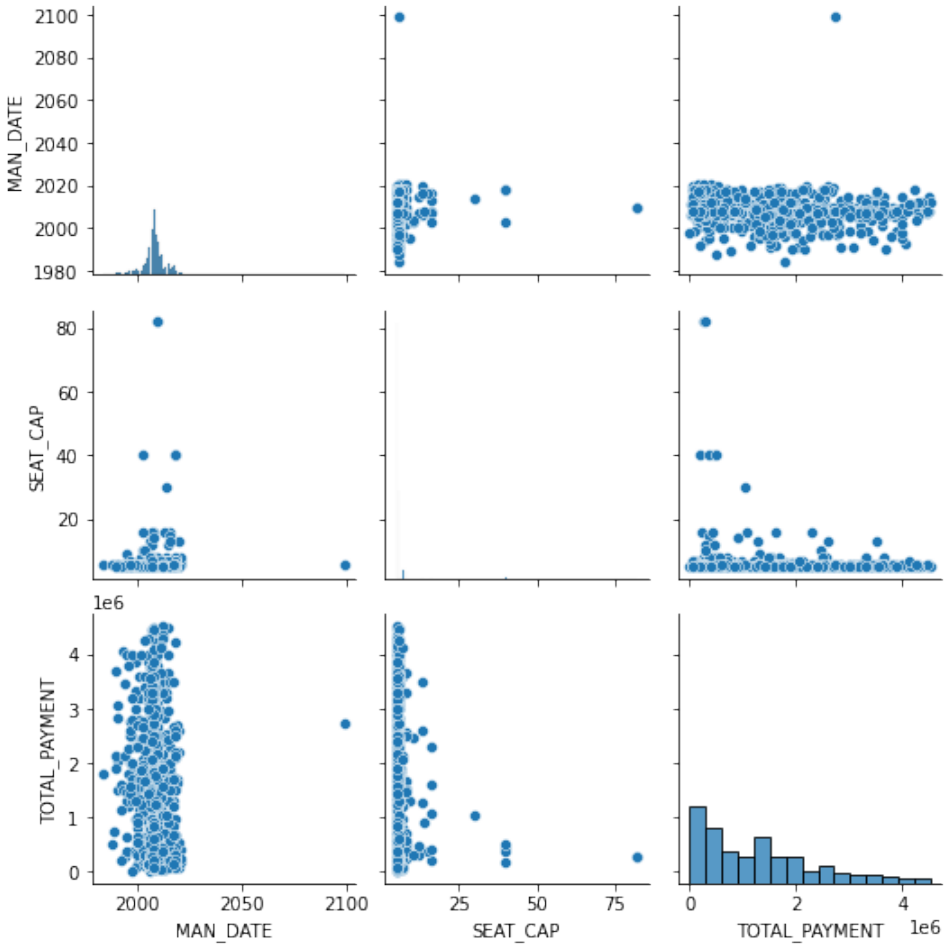
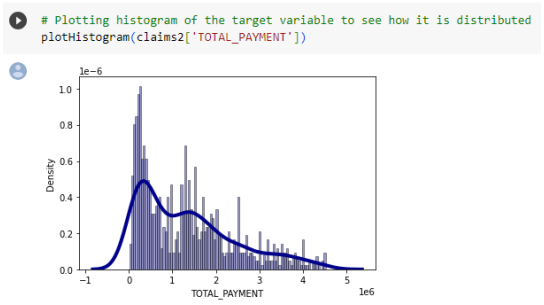


Figure 4.2: Distribution of independent numerical variables against the target variable

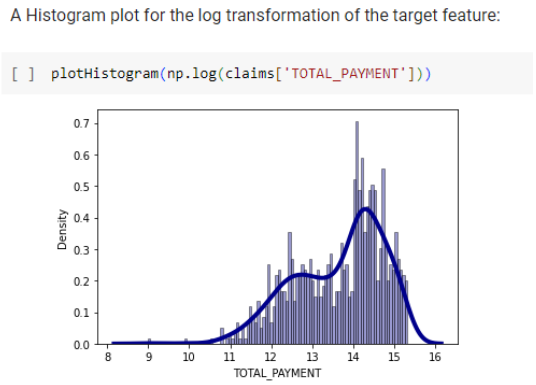
Target Variable Exploration

The target variable of interest which is total claim amount was also plotted to show how the values are distributed using histogram plot. As seen below (4.1), the values are positively skewed which shows that the target variable is not normally distributed. Trying to normalize the target variable by plotting the log-transformed values of the target variable, below are the plots of the before and after log-transformation.

Below are the codes that were run in python to produce the plots in figure (4.1) below.



(a)



(b)

Figure 4.3: Figure showing the distribution of the target variable before and after transformation

Thus, from the plots it can be noted that both the target variable and its log-transformed values are not normally distributed. While the target variable is positively skewed indicating the presence of positive extreme values. These represents the availability of huge claim sizes in auto insurance. Thus, while most of the claims are minor, in auto insurance there is the presence of some huge claims which may have resulted from serious vehicle accidents which mostly lead to loss of lives or serious vehicle damages. In insurance such kinds of damages lead to huge claims as indicated with the long skewed tails on the positive side of the histograms. On the other hand, the log-transformed plot of the target variable shows that it is not normally distributed, however there is an indication of the presence of bi-modal distribution.

Correlation of variables

Figure (4.4) below shows the correlation of variables in the data set. The plot shows which variables are more correlated with the total claim amount. The following command was used in Python to obtain the correlation plots of all the independent variables against the target variable of interest.

```
[ ]  # Plotting heatmap to check for relationships among the variables
     plt.figure(figsize=(15,10))
     sns.heatmap(claims2.corr(),cmap=plt.cm.Reds,annot=True)
     plt.title('Heatmap displaying the relationship between the features of the data',
               fontsize=13)
     plt.show()
```

Variables that have high correlation with the target variable are the ones that will have more predictive power than those that are less correlated. Some independent features such as vehicle damages and personal injuries, vehicles damages and concerned particulars among others appear to be correlated with each other bringing a multicollinearity problem amongst the independent features. However, the bootstrap approach taken by RF ensures that multicollinearity does not affect the performance of the fitted model since RF uses different features that are randomly chosen to form different decision trees. The features' randomization lessens the impact of the variables' multicollinearity.

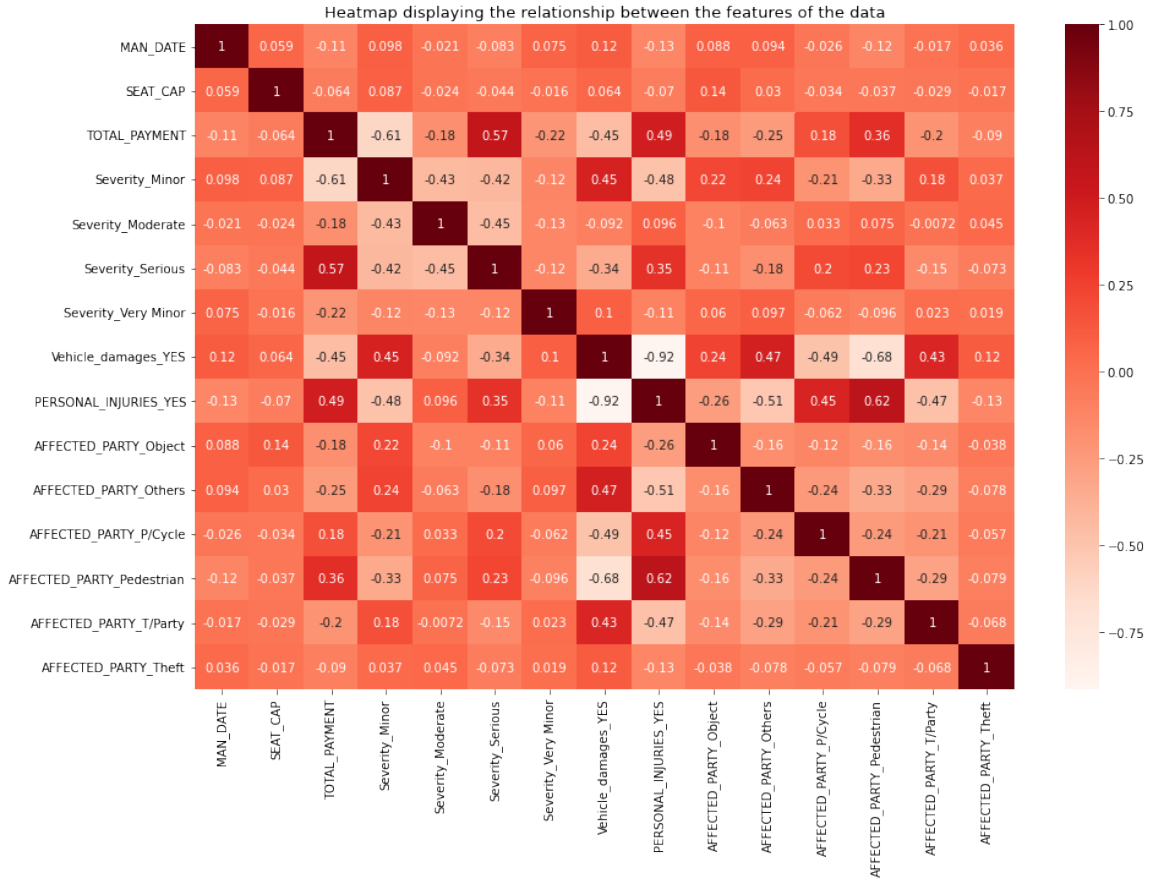


Figure 4.4: Pearson Correlation heat map of Variables

A table below shows the coefficients of correlation between the independent features and the target variable, total claim amount.

Table 4.1: Dataset Variables after extracting other features

Feature	**Correlation**
Severity (Minor)	0.61
Severity (Serious)	0.57
Personal injuries	0.49
Pedestrian	0.36
Pedal cyclists	0.18
SEAT-CAP	-0.064
Theft	-0.09
Manufactured date	-0.11
Object	-0.18
Severity (Moderate)	-0.18
Third party	-0.2
SEVERITY (Very Minor)	-0.22
Others causes	-0.25
Concerned particular (Others)	-0.29
Vehicle damages	-0.45

From the table, it is seen that accident severities such as; minor, serious, and personal injuries are the three variables with positive moderate correlations with the target variable of $0.61, 0.57$ and 0.49, respectively. This implies that there is a positive significant linear relationship between those variables and the claims amount. On the other hand, variables such as vehicle damages and concerned particular (others) have moderate negative relationship with the target feature. This signifies the a slight significant negative linear relationship between such variables and the claims amount. Other variables such as Concerned particular-pedestrian and concerned particular-cylist have weaker positive correlations as seen in the table (4.1) above. Similarly, variables such as SEAT-CAP, Concerned particular-theft, Manufactured date and Concerned particular-third party have weaker negative relationship with the target variable. Overall, variables with weaker positive and negative relationship with the target variable implies that such explanatory variables have less impact on the target variable i.e. the claim amount is less dependent on such explanatory variables.

4.2 Regression Results and Model Evaluation

This part gives the findings of the fitted regression models in predicting the individual claim amount. To achieve this, Random Forest regression, Neural Network regression and Extreme Gradient Boosting (XGB)regression models were used. To evaluate each model's effectiveness in forecasting the target variable, different model assessment measures were applied. The evaluation metrics for each model were compared against each other to establish the best performing model among the three models. To ensure that the models were not over-fitting, the evaluation metrics were examined both for training and test data.

4.2.1 Random Forest Model Results

Random Forest regression model come with a wide range of hyperparameters that needs tuning to ensure that the best model is fitted. To assess the impact of hyperparameter tuning, first a RF model was fitted which acts as a benchmark model to the tuned RF model. Then, RF model was fitted with tuned hyperparameters using RandomizedCv and Grid-SearchCV functions in scikit-learn. Below are the metrics that were obtained from the three models.

Table 4.2: Evaluation metrics for Random Forest Models

MODEL	Train MSE	Test MSE	Train RMSE	Test RMSE	Train MAE	Test MAE	Train R^2	Test R^2
SRF	5.95	17.87	2.44	4.23	1.67	2.94	0.95	0.87
RRF	11.06	15.60	3.33	3.95	2.42	2.80	0.90	0.89
GSRF	11.60	15.28	3.41	3.91	2.49	2.78	0.90	0.89

The results show that the two models, RandomizedCV random Forest (RRF) and Grid-Search Random Forest (GSRF), where parameters were tuned performed better than the Standard Random Forest (SRF) model as can be seen in the table (3.2) above. Before tuning the parameters, the standard RF model had a score of 87% while both the RandomizedCV and GridSearch produced a score of 89% each. In addition to that, training and testing evaluation metrics for the standard RF model differs substantially which is not a good sign as it may imply that the model over-fits to the training data by learning from meaningless noise in the data and fails to achieve the same score on the unseen test data. On the other hand, there are slight differences in the evaluation metrics of the tuned RF models. This means that the two models trains well on

the data and hence producing slightly higher evaluation metrics on the unseen data than their counterpart training metrics. Both the Randomized model and Grid-Search RF model performs better than the standard RF model, with negligible difference. Thus, tuning the parameters has helped the model to improve its efficiency in predicting the target variable.

Random Forest Model Adequacy Analysis

After fitting the models to the data, errors were investigated to find out how they are distributed. Generally, for a model that is well fitted plots of errors needs not to have a trend and must be normally distributed with mean 0 and unknown variance. This means that the scatter plot of errors should have no defined trend and the histogram should be normally distributed with values clustered around zero. Figure (4.2.1) below are the plots of the errors for the fitted RF model.

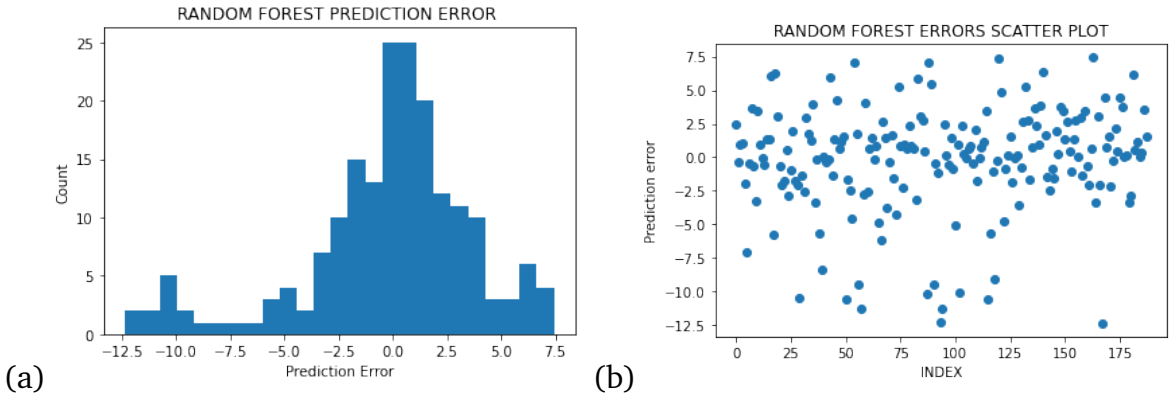


Figure 4.5: Random Forest Error Distribution

Plot of Random Forest Predicted Values

Figure below shows the plot of the predicted against the actual claim amount from the random forest model.The y axis represents the claim values while the x axis represents indexes of the 175 test data points.

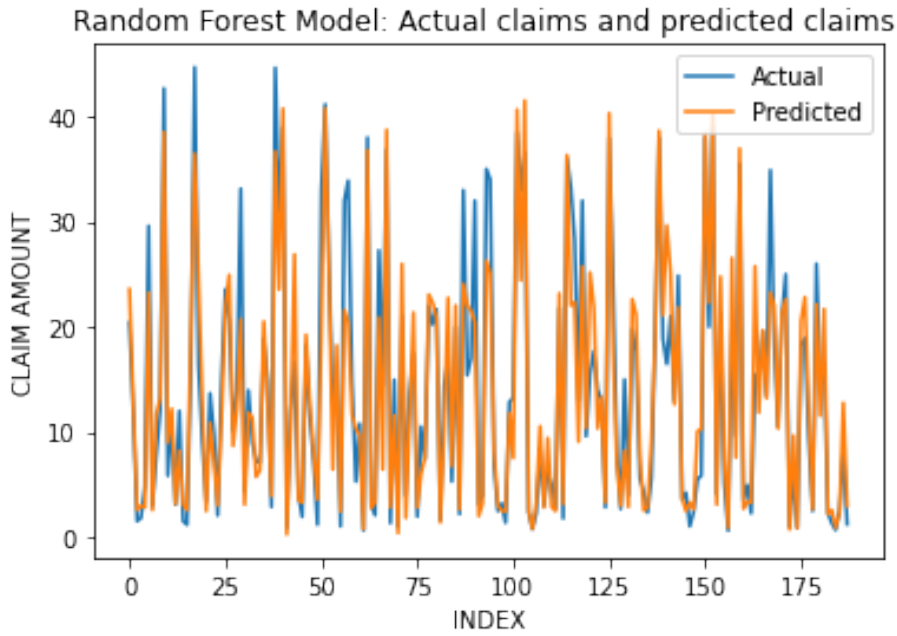


Figure 4.6: Actual claims VS Predicted

4.2.2 Extreme Gradient Boosting Model Results

Another regression results are presented where Extreme Gradient Boosting was used and results are below. Similarly, a standard XGBoost model was fitted before finding the best parameters using the RandomizedCV function and results for the tow fitted models are presented below to assess the impact of hyperparameter tuning. Metrics such as mean squared error, Root mean squared error, mean absolute error and R^2 score are used to compared the performances of the two XGBoost models. Table below shows the evaluation metrics,

Table 4.3: Evaluation metrics for XGBoost Model

MODEL	Train MSE	Test MSE	Train RMSE	Test RMSE	Train MAE	Test MAE	Train R^2	Test R^2
SXGB	5.36	20.53	2.32	4.53	1.44	3.13	0.95	0.85
RXGB	11.49	16.89	3.90	4.11	2.50	2.92	0.90	0.88

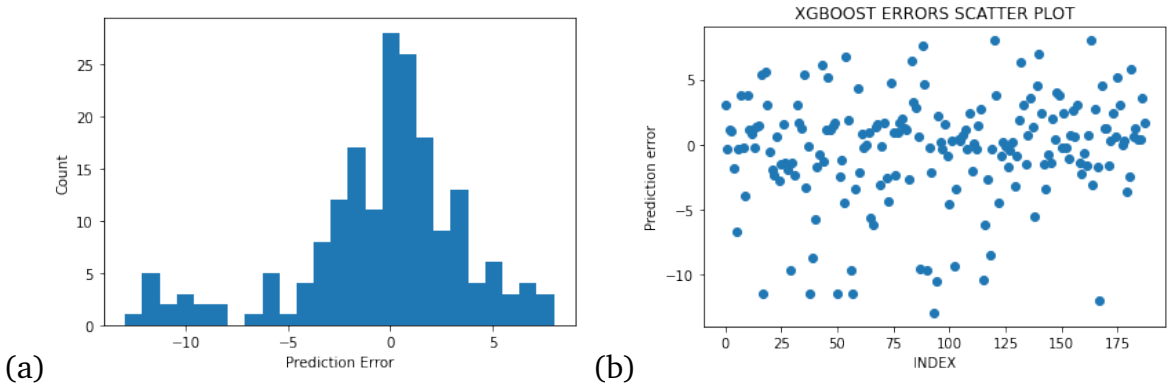
As can be seen from the table (4.3) above, the Randomized XGBoost model out performs the standard XGBoost (SXGB) model. While standard model has test MSE of 20.53, RMSE of 4.53, MAE of 3.13 and R^2 of 85%, the Randomized XGBoost model has the test MSE of 16.89, RMSE

of 4.11, MAE of 2.92 and R^2 of 88%, respectively. This precisely shows that the Randomized XGBoost model better fits the data than the mere standard XGBoost model.

Further to that, the train and test evaluation metrics for the traditional XGBoost model differs considerably and this may mean that the standard XGBoost model over-fits the data as it gives much lesser metrics on the training data than on the testing data. On the contrary, the differences in the evaluation metrics between the training and testing for the RandomizedCV XGBoost model shows that the model neither under-fits nor over-fits the data. This is common for all tree based models, because of the regularization parameters, tree based models reduce the possibility of over-fitting and under-fitting.

Extreme Gradient Boosting Model Adequacy Analysis

In the same way, the errors for the XGBOOST model were analysed to assess how best the model predicts the target variable. This was done by plotting the scatter plot of errors and histogram to check if they satisfy the no trend assumption and the normality assumption.



(a) (b)

Figure 4.7: Extreme Gradient Boosting Error Distribution

Plot of Extreme Gradient Boosting Predicted Values

The predicted values were plotted against the actual values to visualise the fitness of XGBOOST regression model and below is how the plot came out. The y axis represents the claim values while the x axis represents indexes of the 175 test data points.

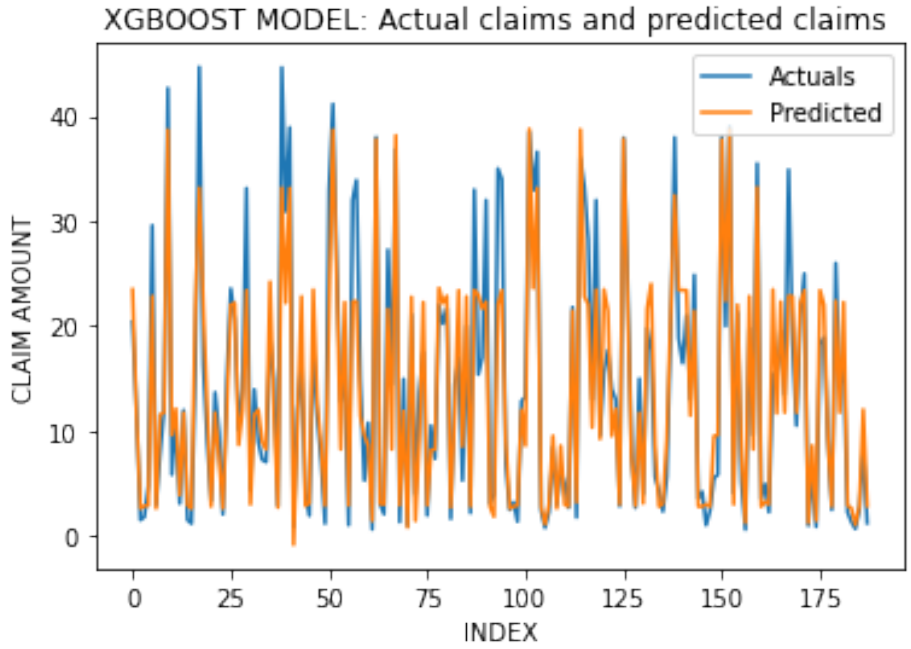


Figure 4.8: Actual claims VS Predicted

4.2.3 Neural Network Model Results

A neural network is one of the advanced ML techniques that that is suitable for non linear relationships that are more complicated to be modelled. Here a neural network was developed using tensor-flow library in python and training and validation was plotted as seen below to see how best the model fits to the training data and consequently, how it performs of new unknown data.

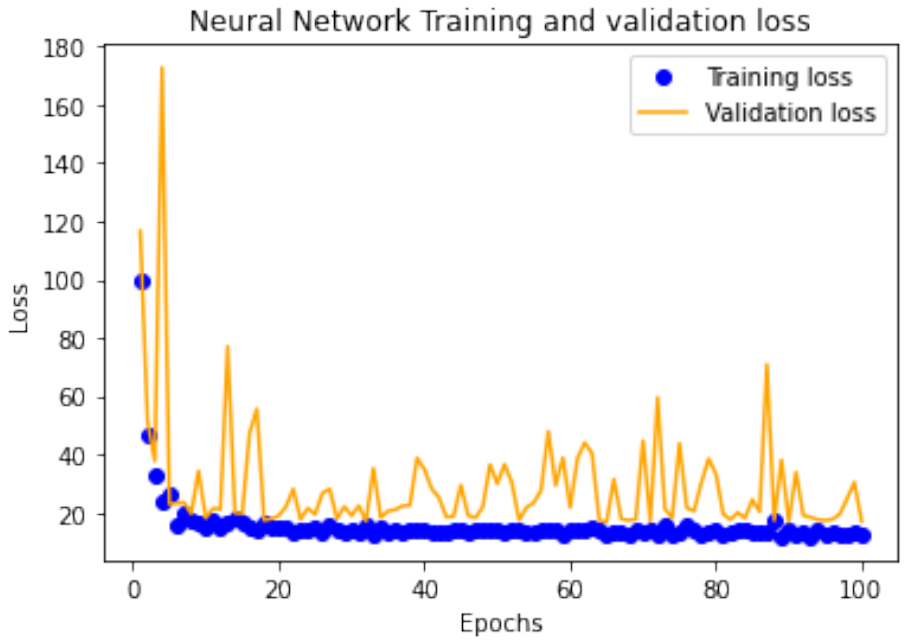


Figure 4.9: Neural Network Training and validation loss

As can be seen in the figure above, the neural network model fits well to the training data as depicted by a smooth converging blue line representing small loss in training. However, the loss in the model on new data is oscillating as can be seen in the un-converging line in yellow which oscillates as the number of epochs increase indicating noise. This indicates that the model generalizes the relationship between the predictor variables and the target variable better on the training dataset but fails to do the same on the unknown test dataset. However, the rate of convergence increases drastically with increase in epochs, and thereby decreasing both training loss as well as validation loss. Below is a plot of the training model and the validation model.

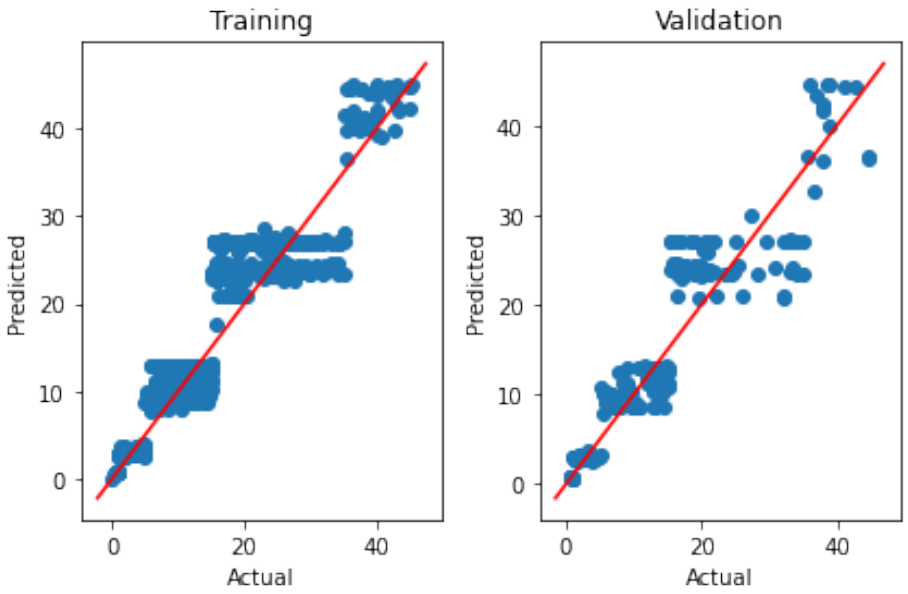


Figure 4.10: Neural Network Training and validation loss

The plot (4.2.3) above shows that the neural network regression model generalizes the data well. Evaluation metrics were computed and below is a table representing the performance of the neural network regression model.

Table 4.4: Evaluation metrics for Random Forest Models

MODEL	Train MSE	Test MSE	Train RMSE	Test RMSE	Train MAE	Test MAE	Train R^2	Test R^2
NN	12.01	17.06	3.47	4.13	2.50	2.91	0.90	0.88

Neural network Model Adequacy Analysis

Below are the scatter plot and distribution plot of errors of the neural network regression model.

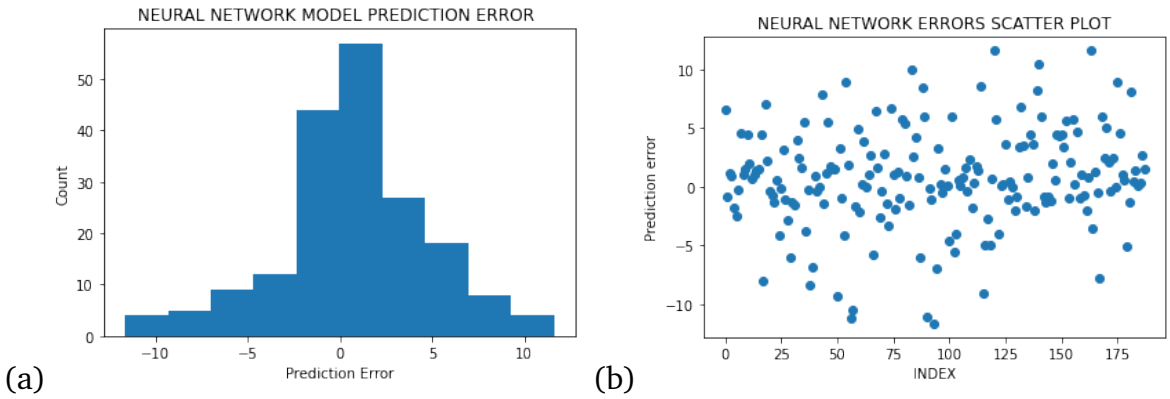


Figure 4.11: Neural Network Model Error Distribution

Both plots shows that the neural network model fits the data well as both the normality assumption of errors is satisfied and the scatter plot has no trend.

The plot of the actual against the predicted values is seen below where the y axis represents the claim values while the x axis represents indexes of the 175 test data points.

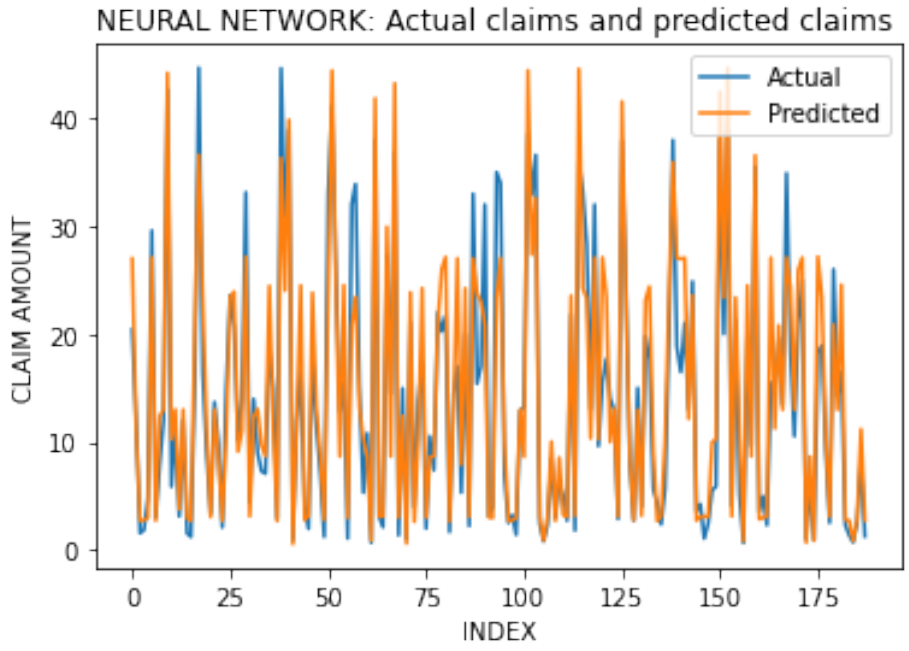


Figure 4.12: Neural Network Model Actual VS Predicted

Feature Importances

Tree based models in machine learning have the ability of identifying features that were important in predicting the response variables, in particular, Random Forests and Extreme Gradient Boosting models allow one to see which features are important in predictions. Consequently, feature importances were plotted in python for the two models and below are the plots for the models.

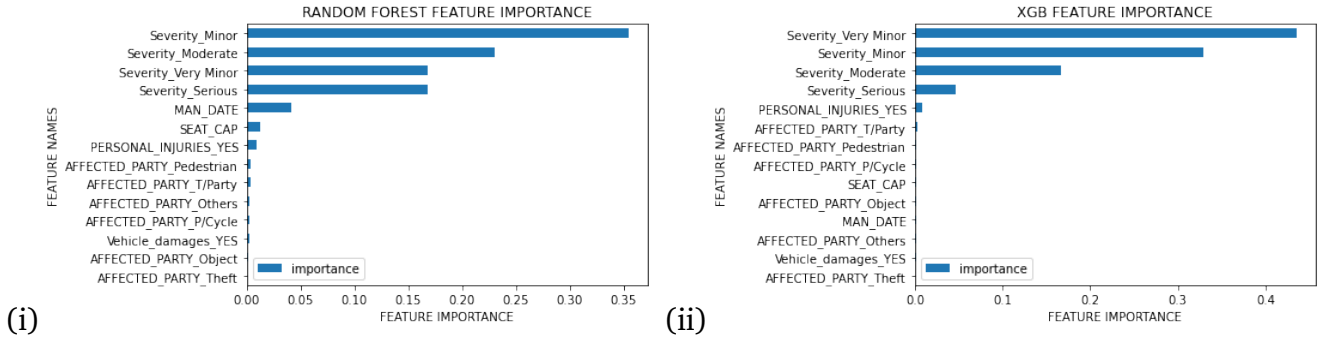


Figure 4.13: Random Forest and XGBoost Feature Importances

A total of 14 features were used to achieve the objectives of the study, and as can be seen from the plot above (4.2.3), not all features were important in predicting the response variable. Some features contributed more as indicated by the length of the bars. From the RF model, it is seen that severities; Minor, Moderate, very minor,serious and the manufactured date are the to five important features. Where as in XGBoost model gives features as severities; very minor, minor, moderate, serious and personal injuries as the top five contributing variables.

4.2.4 Model Results Comparisons

Here we compare the performances of the three models fitted to together by looking at the evaluation metrics. Overall, Random Forest (GridSearch) and Extreme Gradient Boosting (Randomized) model slightly outperforms the Neural Network model as can be seen from the evaluation metrics. Random Forest almost performs the same as Extreme Gradient Boosting model as the evaluation metrics for the two models are approximately the same. This tells that the tree based models produces the same prediction results.

Table 4.5: Model Results Comparisons

MODEL	Train MSE	Test MSE	Train RMSE	Test RMSE	Train MAE	Test MAE	Train R^2	Test R^2
RF	11.60	15.28	3.41	3.91	2.49	2.78	0.90	0.89
XGB	11.49	16.89	3.90	4.11	2.50	2.92	0.90	0.88
NN	12.01	17.06	3.47	4.13	2.50	2.91	0.90	0.88

The RF model achieves a testing score of 89% whereas the XGBoost achieves a testing score of 88% with NN models achieving a testing score of 88% as well. As can be seen from the table above, there is slight difference in evaluation metrics on training and testing data. This tells that the models do not over-fit or under-fit. Overall, all the models produce interesting results and hence using any of the models in prediction of individual claim reserves can ensure efficiency results.

4.3 DISCUSSIONS AND CONCLUSIONS

4.3.1 Discussions

The primary purpose of the study was to predict individual loss reserves in auto insurance. Using machine learning regression techniques such as Random Forest, Extreme Gradient Boosting and Neural Networks, models were fitted on actual insurance claims data for closed claim cases. These techniques are robust and comes with parameters for optimization to achieve highest model performance without over-fitting and under-fitting. Two sets of data were used; one set (training data set) for model training and the other set (testing dataset) for model valuation. Machine learning models being models that learn from experience, were used to capture the patterns of different individual claim cases while associating the claim amounts to individual risk factors. The study results has shown that Random Forest model achieved an accuracy score of 89% while Extreme gradient Boosting and Neural Networks model both achieved accuracy scores of 88%.

The analysis in the study has shown that risk factors related to the accident details are more influential in affecting the amount an individual may claim, and consequently a possible reserve

amount that should be set aside for that particular claim. This means severity of the damage, personal injuries, vehicle damages among other risk factors, have more impact on the amount of claims. This agrees with theory Bortoluzzo et al. (2011) as insurance claim amounts tend to increase with increase in accident severities. On the other hand, claim amounts are small when there are no serious injuries registered during the accident. Further to this, a claim amount tend to be huge when there were casualties registered than when only the vehicles were damaged. However, when the vehicle damages are serious i.e. the vehicle overturned, the claim amount is also big. Thus, individual claim amounts increases with increase in severity of the accident and the severities in body injuries registered.

In agreement with Wuthrich Wüthrich (2018) and baudry Baudry and Robert (2019), the study has shown that utilizing individual claims data and machine learning techniques, it is possible to calculate individual claims reserves. Although, Wuthrich focused on simulated data, this study has shown that the results that were obtained on simulated data can also be achieved on actual insurance data as the methods achieved highest accuracy scores. Unlike other similar researches Bornhuetter and Ferguson (1972) Mack (1993) Arico et al. (1972) where traditional methods were used to predict claims as an aggregate value, this study has provided a more flexible methods for which each claim case can be analysed and be looked at, individually.

Further, in his recent research Dong Qiu Qiu (2019) worked on the use of ML in loss reserving where the outcomes of applying ML techniques like Neural Networks (NN) and Random Forest (RF) were then contrasted with those of using traditional techniques on both simulated and actual data. However, Dong failed to base his conclusions on real individual data as his data was based on aggregation. Thus, he failed to compare the effectiveness of the machine learning methods to the traditional and stochastic approaches, and determine which approach best suited whatever scenario. This study not only has it provided the efficiency of the models used, but it has also shown that Random Forest achieves the highest accuracy score in prediction of individual claims insurance. Thus, this study has shown that machine learning models are capable of producing reliable and effective results in claims predictions and therefore, they can be adopted in insurance industry. Finally, although the data used in this study came from one auto insurance company in Malawi, it should however be noted that the findings of the study

can be extended to any auto insurance company in Malawi, and the rest of the companies across the world provided the variables are the same.

4.3.2 Conclusions

Claim reserving in insurance companies is of paramount importance for the smooth operations of business in insurance industry. Insurers need to make sure that at each particular time, they have enough claims reserves to cover the damages of their policyholders. Having available enough loss reserves will ensure that the insurer covers the losses of its clients on time. This will impact the operations of the company positively as it will help to ensure customer loyalty and will increase customer retention. Further to that, accurate individual claim reserves helps insurers to avoid liquidation of their assets that may have been invested in other financial securities. It is therefore imperative that insurance companies maintain accurate reserve predictions at any particular moment in time. This will help them to cover the losses that their clients may have encountered without any of their business operations being disturbed.

The findings of the study have shown remarkable results by considering the evaluation metrics of the fitted models. The results shows that Random Forest model outperforms the other two models namely, Extreme Gradient Boosting and Neural Networks. The RF produces the lowest error values compared to the XGBoost and Neural Network model as can be seen in the table (4.3) above. The second best model is the XGBoost model and followed by the Neural Network model. The RF model produces an accuracy score of 89% which is the highest score among the scores of the fitted models, XGBoost comes on second producing an accuracy score of 88% while Neural Network model attains a score of 88% as well. The robustness and flexibilities of these machine learning regression models ensures reliable and efficient results are produced. The evaluation metrics achieved by the models guarantees their usability in predicting the target variable given the necessary predicting variables.

Further, RF achieves Mean Absolute Error of 2.78, XGBoost achieves Mean Absolute Error of 2.92 and Neural Network attains a Mean Absolute Error of 3.15. This shows that the RF model performs better in predictions of individual claim reserves followed by the XGBoost model. Gen-

erally, tree based models are more stable than their machine learning regression counterparts as depicted by the findings. This ensures that the models are very reliable of producing desirable predictions. It is therefore very important that insurance companies embrace these analytics in their operations for data-driven insights in their daily operations.

4.3.3 Future Works

Generally, machine learning models require huge amount of data to better generalize the relationship, in particular, Neural Networks models requires a lot of instances to train the model on. This allows the models to learn the patterns very well and produce better predictions on the unseen data. Future reseraches may consider getting data from various companies for better generalization.

Additionally, there are various machine learning models for regression however, only Random Forest, Extreme gradient boosting and Neural networks were explored. The choice came about because of their stability in fitting models to non linear relationships. As such some ML models could also be explored and compare the findings with the used models to see which models fits best to address the individual loss reserving issue. Thus, future researchers should explore other ML methods and see their performances against the used methods.

Further, during the model optimization where Grid search and Randomization search were used to find the best hyper-parameters, only a few parameters were optimised due to computing challenge and speed. However, machine learning models have a variety of hyper-parameters that needs to be adjusted to increase the model accuracy. It will therefore be great to further increase the number of these parameters in the future.

Finally, the study used ensemble learning where the averages of the final out was taken from the decisions of the same model. A more hybrid ensemble learning can be considered in the future where the final predicting model should be the combination of multiple machine learning regression models.

References

Aldrich, C. (2020). Process variable importance analysis by use of random forests in a shapley regression framework. *Minerals*, 10(5):420.

Arico, N., Bloom, L., Hillebrandt, A., Horowitz, B., Lange, D., Moore, K., Nation, D., and Patel, C. (1972). Bornhuetter-ferguson initial expected loss ratio working party paper.

Baudry, M. and Robert, C. Y. (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, 35(5):1127–1155.

Björkwall, S. (2011). *Stochastic claims reserving in non-life insurance: Bootstrap and smoothing models*. PhD thesis, Department of Mathematics, Stockholm University.

Bornhuetter, R. L. and Ferguson, R. E. (1972). The actuary and ibnr. In *Proceedings of the casualty actuarial society*, volume 59, pages 181–195.

Bortoluzzo, A. B., Claro, D. P., Caetano, M. A. L., and Artes, R. (2011). Estimating total claim size in the auto insurance industry: a comparison between tweedie and zero-adjusted inverse gaussian distribution. *BAR-Brazilian Administration Review*, 8:37–47.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brownlee, J. (2016). *XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn*. Machine Learning Mastery.

Brownlee, J. (2021). Regression metrics for machine learning. https://machinelearningmastery.com/regression-metrics-for-machine-learning/. Accessed: 2022-06-23.

Carrato, A. and Visintin, M. (2019). From the chain ladder to individual claims reserving using machine learning techniques. In *ASTIN Colloquium, Cape Town, South Africa, April*, pages 2–5.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

England, P. D. and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, 8(3):443–518.

Gabrielli, A. (2021). An individual claims reserving model for reported claims. *European Actuarial Journal*, 11(2):541–577.

Grace, M. F. and Leverty, J. T. (2012). Property–liability insurer reserve error: Motive, manipulation, or mistake. *Journal of Risk and Insurance*, 79(2):351–380.

Gründl, H., Gal, J., et al. (2017). The evolution of insurer portfolio investment strategies for long-term investing. *OECD Journal: Financial Market Trends*, 2016(2):1–55.

Ha, V.-S. and Nguyen, H.-N. (2016). Credit scoring with a feature selection approach based deep learning. In *MATEC Web of Conferences*, volume 54, page 05004. EDP Sciences.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer.

Hindley, D. (2017). *Claims Reserving in General Insurance*. Cambridge University Press.

Kennedy, K. (2013). Credit scoring using machine learning.

Kumar, K. and Rao, A. C. S. (2018). Breast cancer classification of image using convolutional neural network. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*, pages 1–6. IEEE.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin: The Journal of the IAA*, 23(2):213–225.

Mahlow, N. and Wagner, J. (2016). Process landscape and efficiency in non-life insurance claims management: An industry benchmark. *The Journal of Risk Finance*, 17(2):218–244.

Mitchell, T. M. and Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.

Patel, A. A. (2019). *Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data*. O'Reilly Media.

Qiu, D. (2019). *Individual Claims Reserving: Using Machine Learning Methods*. PhD thesis, Concordia University.

Raeva, E. and Pavlov, V. (2017). Planning outstanding reserves in general insurance. In *AIP Conference Proceedings*, volume 1895, page 050009. AIP Publishing LLC.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Samuel, A. L. (1959). Machine learning. *The Technology Review*, 62(1):42–45.

Schmidt, K. D., Zocher, M., et al. (2007). *The bornhuetter-ferguson principle*. Techn. Univ., Inst. für Mathematische Stochastik.

Seif, G. (2019). Understanding the 3 most common loss functions for machine learning regression. https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learning-regression. Accessed: 2022-06-24.

Sober, E. (1981). The principle of parsimony. *The British Journal for the Philosophy of Science*, 32(2):145–156.

Straub, E. and Grubbs, D. (1998). The faculty and institute of actuaries claims reserving manual. volume 1 and 2. *ASTIN Bulletin: The Journal of the IAA*, 28(2):287–289.

Taylor, G., McGuire, G., and Sullivan, J. (2008). Individual claim loss reserving conditioned by case estimates. *Annals of Actuarial Science*, 3(1-2):215–256.

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, 2018(6):465–480.

APPENDIX

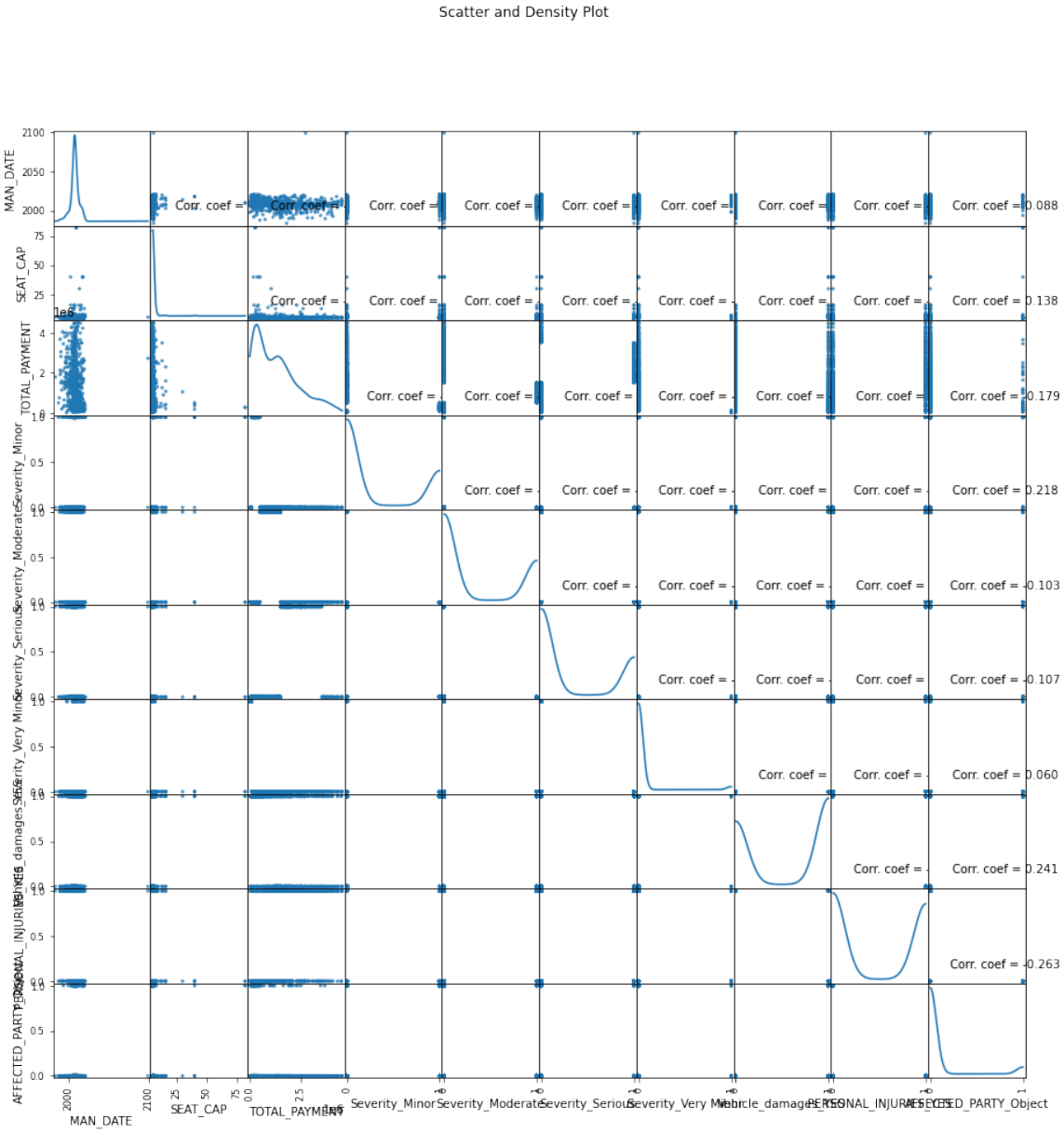


Figure 14: Scatter plot matrix of the variables after one hot encoding

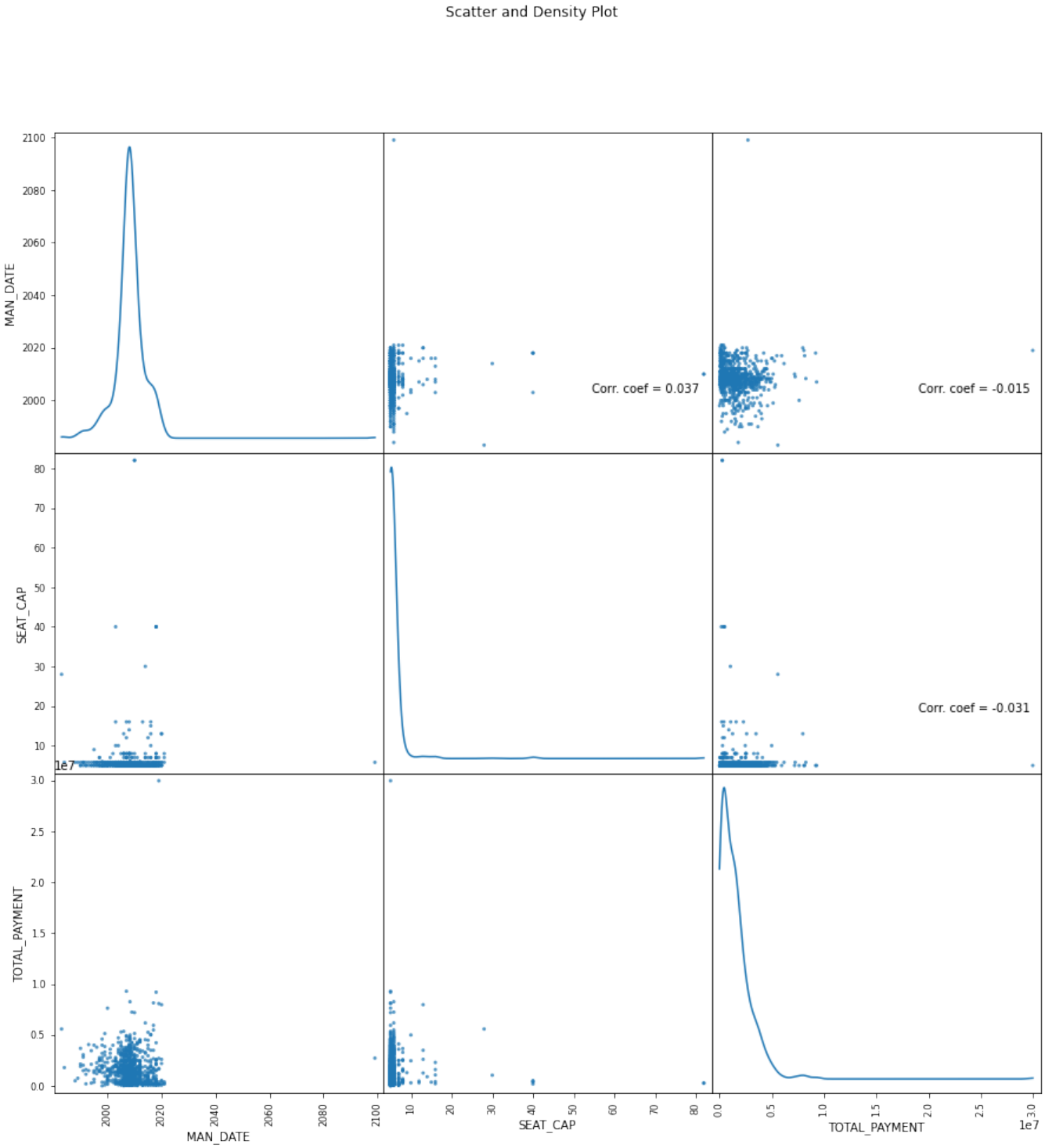
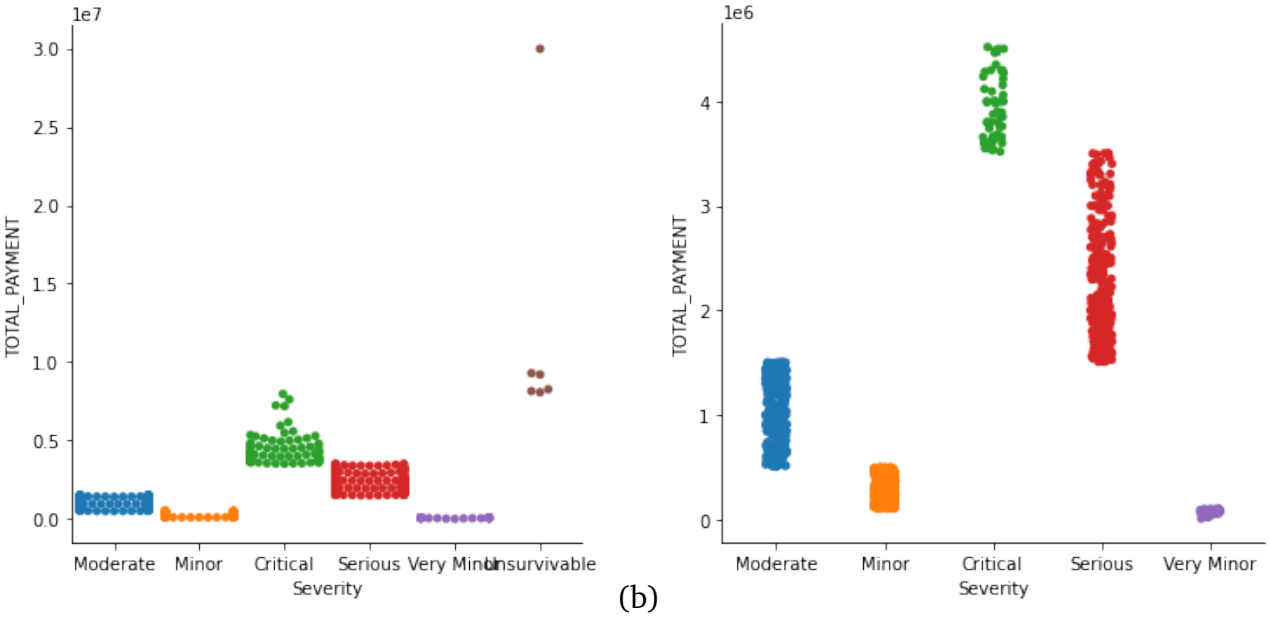


Figure 15: Scatter plot matrix for continuous variables i.e. Man-Date, Total claim amount and Manufactured date



(a) (b)

Figure 16: Catplot showing the relationship between the claim amount and severity categories, before and after removal of outliers

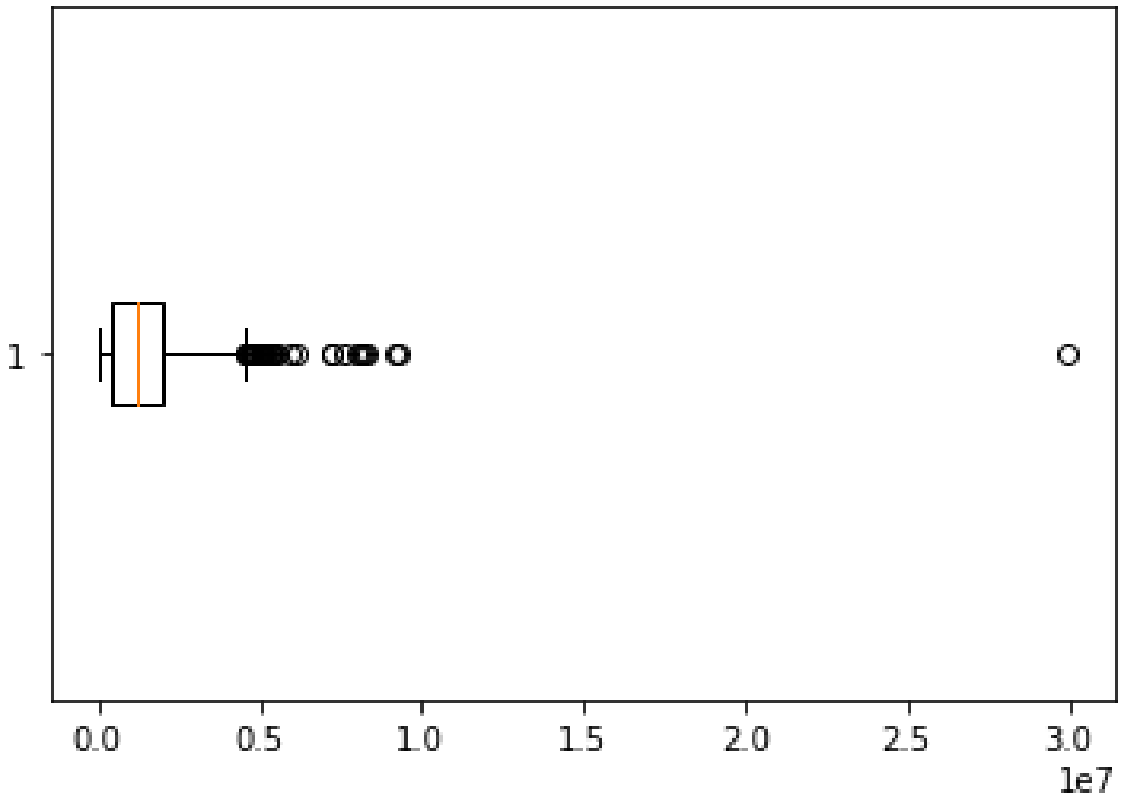


Figure 17: Box plot showing the distribution of the target variable (Claim amount) before outliers were removed

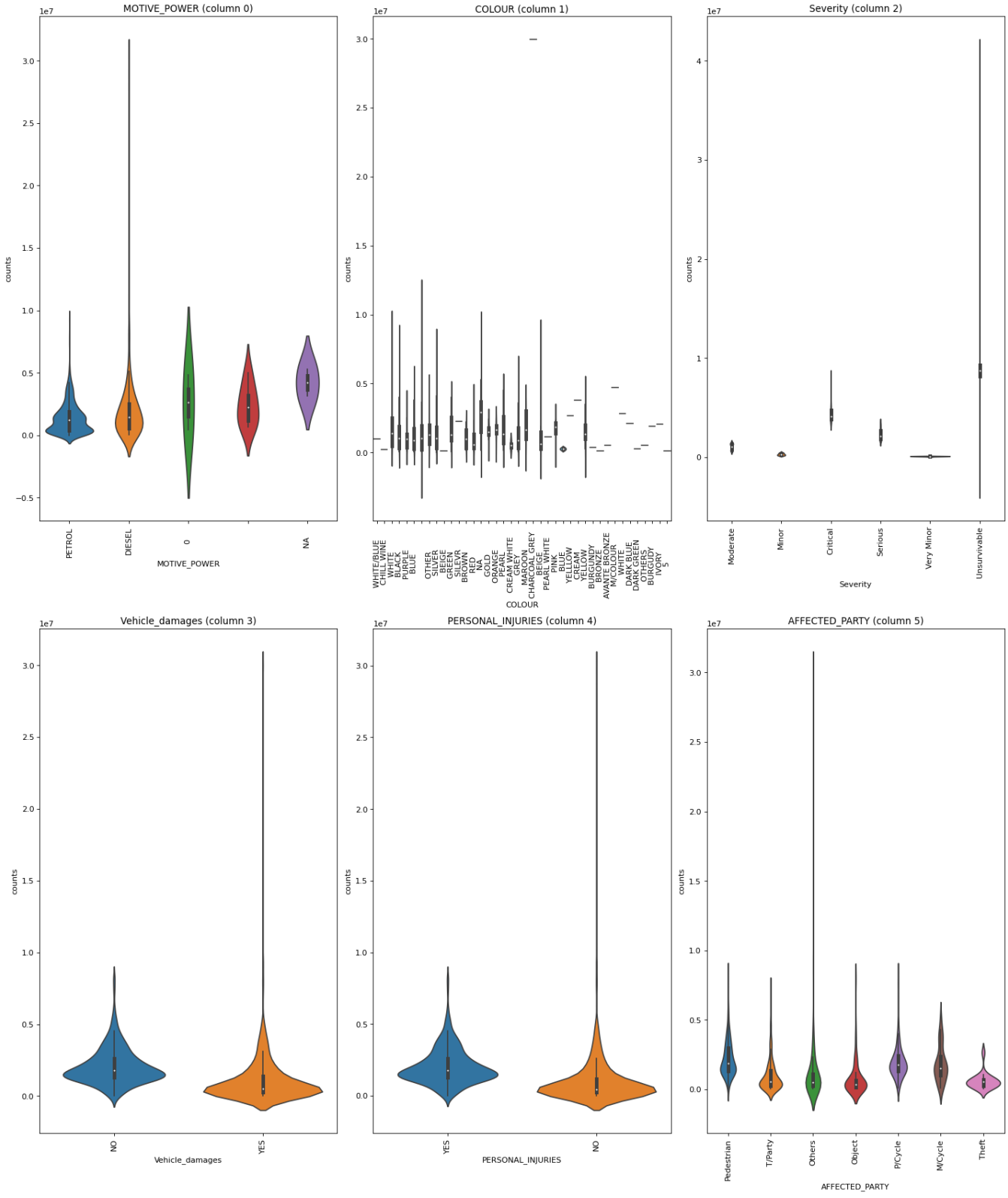


Figure 18: Violin plot showing the distributions of the variables before removal of outliers

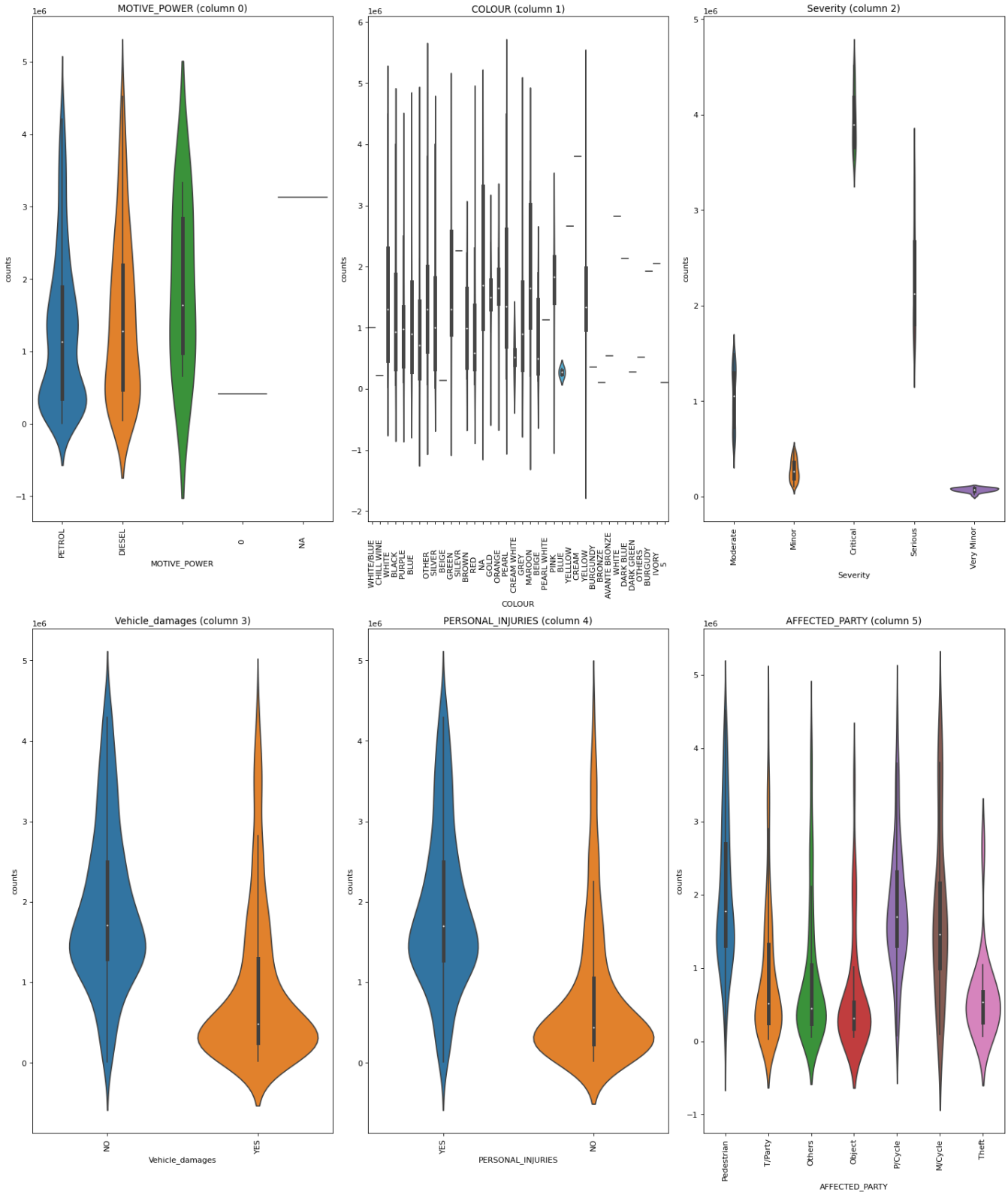


Figure 19: Violin plot showing the distributions of the variables after removal of outliers

djbeyamu@gmail.com.pdf

ORIGINALITY REPORT

17%	13%	8%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	1%
2	Submitted to University of Malawi - The Polytechnic Student Paper	1%
3	tel.archives-ouvertes.fr Internet Source	1%
4	orcid.org Internet Source	1%
5	Submitted to RMIT University Student Paper	<1%
6	www.researchgate.net Internet Source	<1%
7	Submitted to Kang Chiao International School East China Campus Student Paper	<1%
8	Submitted to University of Strathclyde Student Paper	<1%
9	erepository.uonbi.ac.ke	

	Internet Source	<1%
10	towardsdatascience.com Internet Source	<1%
11	Akshara Makrariya, Shishir Kumar Shandilya, Smita Shandilya, Michal Gregus, Ivan Izonin, Rahul Makrariya. "Mathematical Simulation of Behavior of Female Breast consisting Malignant Tumor during Hormonal changes", IEEE Access, 2022 Publication	<1%
12	corpus.ulaval.ca Internet Source	<1%
13	onlinelibrary.wiley.com Internet Source	<1%
14	par.nsf.gov Internet Source	<1%
15	Jakob Karolus, Paweł W. Wozniak, Lewis L. Chuang, Albrecht Schmidt. "Robust Gaze Features for Enabling Language Proficiency Awareness", Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17, 2017 Publication	<1%
16	Submitted to University College London Student Paper	<1%

17	Submitted to University of Ulster Student Paper	<1 %
18	www.casact.org Internet Source	<1 %
19	bip.pwr.edu.pl Internet Source	<1 %
20	Submitted to Faculty of Commerce, Cairo University Student Paper	<1 %
21	scholar.mzumbe.ac.tz Internet Source	<1 %
22	www.ressources-actuarielles.net Internet Source	<1 %
23	Submitted to Loughborough University Student Paper	<1 %
24	Submitted to Queen's University of Belfast Student Paper	<1 %
25	Submitted to University of Hertfordshire Student Paper	<1 %
26	Submitted to University of Southampton Student Paper	<1 %
27	www.mdpi.com Internet Source	<1 %