University of Ghana http://ugspace.ug.edu.gh

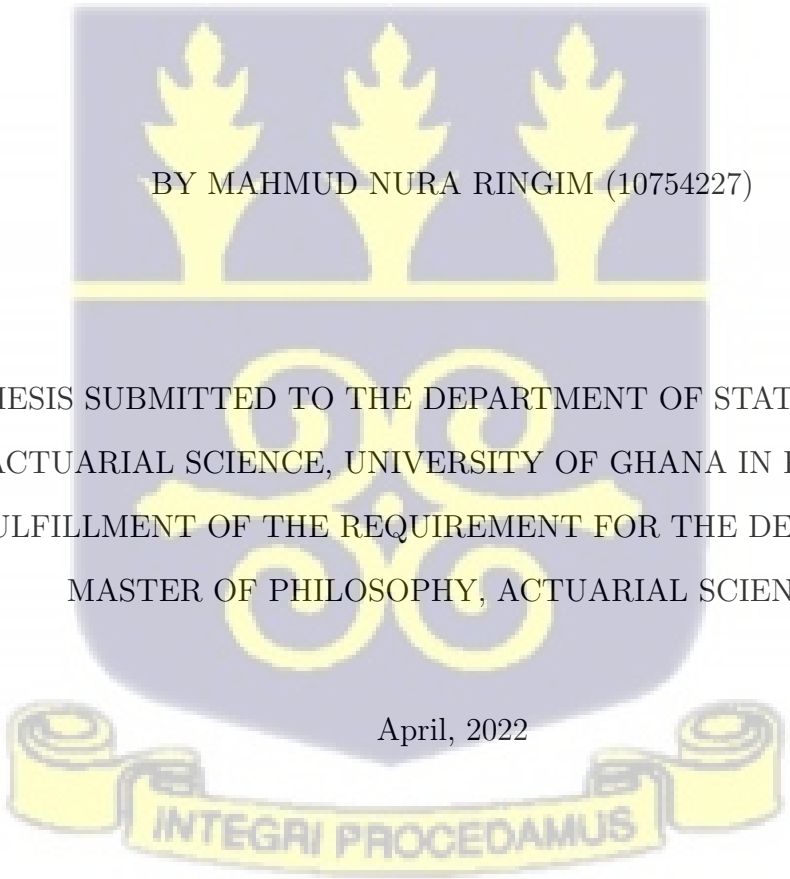
UNIVERSITY OF GHANA, LEGON

AN INVESTIGATION INTO MODELING NON-LIFE INSURANCE CLAIMS IN THE NIGERIAN AND GHANAIAN INSURANCE MARKET

BY MAHMUD NURA RINGIM (10754227)

A THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, UNIVERSITY OF GHANA IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF PHILOSOPHY, ACTUARIAL SCIENCE

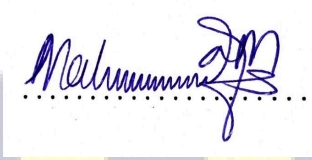
April, 2022

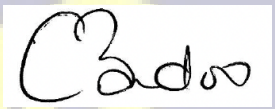


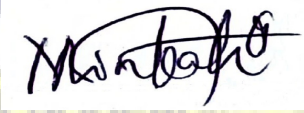
University of Ghana http://ugspace.ug.edu.gh

Declaration

I hereby declare that this submission is my own work towards the award of the Master of Philosophy degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

Mahmud Nura Ringim		21/04/2022
Student		
(10754227)	Signature	Date

Certified by:		
Dr. Isaac Baidoo		28/4/2022
Principal Supervisor	Signature	Date

Certified by:		
Dr. Richard Minkah		30/04/2022
Co-Supervisor	Signature	Date

University of Ghana http://ugspace.ug.edu.gh

Dedication

To my family

University of Ghana http://ugspace.ug.edu.gh

Abstract

For calculating non-life insurance premiums, actuaries rely on separate modeling of frequency and severity using covariates to explain the claims loss exposure. In this thesis, we focus on the insurance claims severity amount. Two separate insurance claims data were analyzed using some selected Tree-Based Machine Learning (ML) Algorithms namely; the Classification and Regression Tree (CART), Random Forest (RF), and Gradient Boosting (GB) Models. The predictive performance of the selected models were compared using the Coefficient of determination (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). In the application of the selected models, this Thesis relied on two different insurance claims data; The Nigerian and the Ghanaian Insurance claims dataset. The Nigerian dataset had 10,017 observations from paid claims with 4 explanatory variables, while the Ghanaian dataset had 5,495 observations with 7 explanatory variables. In the analysis, 70% of the data were used for training and 30% for testing. Both datasets were compared in terms of the selected performance measures. The results show that the Random Forest model of the claims amount had the overall best performance for both Nigerian and Ghanaian Dataset.

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

Acknowledgment

All praise to the Almighty, the Beneficent, the Merciful. I thank Allah for giving me the wisdom to successfully complete the Actuarial Science Programme. I would like to appreciate my sponsor, the Tertiary Education Trust Fund (TETFUND) for giving me the opportunity to acquire much knowledge and experience. I would also like to thank our former Vice-Chancellor (FUD) Prof. Fatima Batul Mukhtar and our Current Vice-Chancellor (FUD) Prof. Abdulkarim Sabo Muhammad, our former Dean (FMS-FUD) Ass. Prof. Sunusi Sa'ad Ahmad and our current Dean (FMS-FUD) Dr. Abdusslam Mas'ud, and our former HOD (FUD) Dr. Fatima Abdulkarim Muhammad and our current HOD (FUD) Dr. Aminu Ma'aruuf Nass for their endless support and guidance.

My sincere appreciation goes to my Principal Supervisor Dr. Isaac Baidoo for the continuous guidance, support, understanding, patience, encouragement, wisdom and vast knowledge. Dr. Baidoo has encouraged and supported me by sacrificing the time to listen to my concerns and challenges during the course of my Masters programme. Thank you Sir. I am also grateful to my Co-Supervisor, Dr. Richard Minkah, for his advice and assistance which helped me in diverse ways.

A Special thanks goes to our Head of Department in person of Prof. K. Doku-Amponsah, and all the Lecturers in the Department of Statistics and Actuarial Science, University of Ghana for their enlightening, support, encouragement and teaching. I am also grateful to all my class mates. Thank you all.

I would like to also express my infinite appreciation and gratitude to my parents, Alhaji Nura Salisu Ringim and Haj Hadiza Abubakar Mai-Sugar Ringim, for their unending love, and guidance in my life. May Allah Reward you with best in this world and the hereafter.

University of Ghana http://ugspace.ug.edu.gh

Contents

Declaration v

Dedication v

Abstract v

Acknowledgement v

Abbreviation vii

List of Abbreviation viii

List of Tables x

List of Figures xii

1 Introduction 1

1.1 Background of the Study 3

1.2 Problem Statement 7

1.3 Objectives of Study 8

1.4 Research Questions 9

1.5 Scope and Limitations 9

1.6 Significance of Study 10

1.7 Organization of Study 10

2 Literature Review 11

2.1 Decision Tree 11

2.2 Theoretical Background of the Tree-Based Models 13

University of Ghana http://ugspace.ug.edu.gh

2.2.1 The Concept of Tree-Based Model Algorithms 14
2.2.2 Classification and Regression Tree (CART) 15
2.2.3 Bagging and Random Forests 16
2.2.4 Gradient Boosted Regression Trees 17
2.3 Empirical Application of the Tree-Based Machine Learning Models 18
2.3.1 Insurance Claims Modeling in the Nigerian Insurance Market 25
2.3.2 Historical Development of Insurance in Nigeria 28
2.3.3 Brief Overview of the Insurance Industry 29
2.3.4 Classifications of Insurance Services 30

3 Methodology 32
3.1 Overview of the Data 32
3.1.1 Empirical Models 33
3.1.2 Classification and Decision Tree (CART) 34
3.1.3 Random Forest 36
3.1.4 Gradient Boosting 37
3.2 Tuning Parameter Method 38
3.2.1 Grid Search 39
3.2.2 Cross-Validation 40
3.3 Performance Measures of the Tree-Based Models 41
3.3.1 Coefficient of determination (R^2) 42
3.3.2 Mean Absolute Error (MAE) 42
3.3.3 Mean Squared Error (MSE) 43
3.4 Statistical Computing Environment 43

4 Data Analysis and Discussion 44
4.1 Data Description 44
4.2 Descriptive Statistics 44
4.3 Fitting the Machine Learning Models: Nigerian Data 47
4.3.1 The Classification and Regression Tree (CART) 47

University of Ghana http://ugspace.ug.edu.gh

4.3.2 Random Forest 51
4.3.3 Gradient Boosting 54
4.4 The Ghanaian Dataset 58
4.4.1 Description of Data 58
4.4.2 Descriptive Statistics 58
4.5 Fitting the Machine Learning Models: The Ghanaian Dataset . . 59
4.5.1 The Classification and Regression Tree (CART) 59
4.5.2 Random Forest 62
4.5.3 Gradient Boosting 65
4.6 Performance Measures of the Machine Learning Models 69

5 Summary, Conclusion and Recommendation 72
5.1 Summary of Findings 72
5.1.1 The Nigerian Insurance Data 72
5.1.2 The Ghanaian Insurance Data 73
5.2 Conclusion and Recommendation 74

References 77

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

List of Abbreviation

AID ..... Automatic Interaction Detection

AUCROC ..... Area under the Curve Receiver Operating Characteristics

CART ..... Classification and Regression Tree

CRUISE ..... Classification Rule with Unbiased Interaction Selection and Estimation

CP ..... Complexity Parameter

DB ..... Delta Boosting

DT ..... Decision Tree

EVT ..... Extreme Value Theory

FACT ..... Fast and Accurate Classification Tree

FRA ..... Federal Railroad Administration

RF ..... Random Forest

GLM ..... Generalized Linear Models

GB ..... Gradient Boosting

GH ..... Ghana

GUIDE ..... Generalized, Unbiased, Interaction Detection and Estimation

ID3 ..... Iterative Dichotomiser 3

ML ..... Machine Learning

MAE ..... Mean Absolute Error

University of Ghana http://ugspace.ug.edu.gh

MLF ..... Multi-Level Factors

NAICOM ..... National Insurance Commission

NB-GPD ..... Negative Binomial-Generalized Pareto Distribution

NG ..... Nigeria

NIA ..... Nigeria Insurer Association

OOB ..... Out of Bag

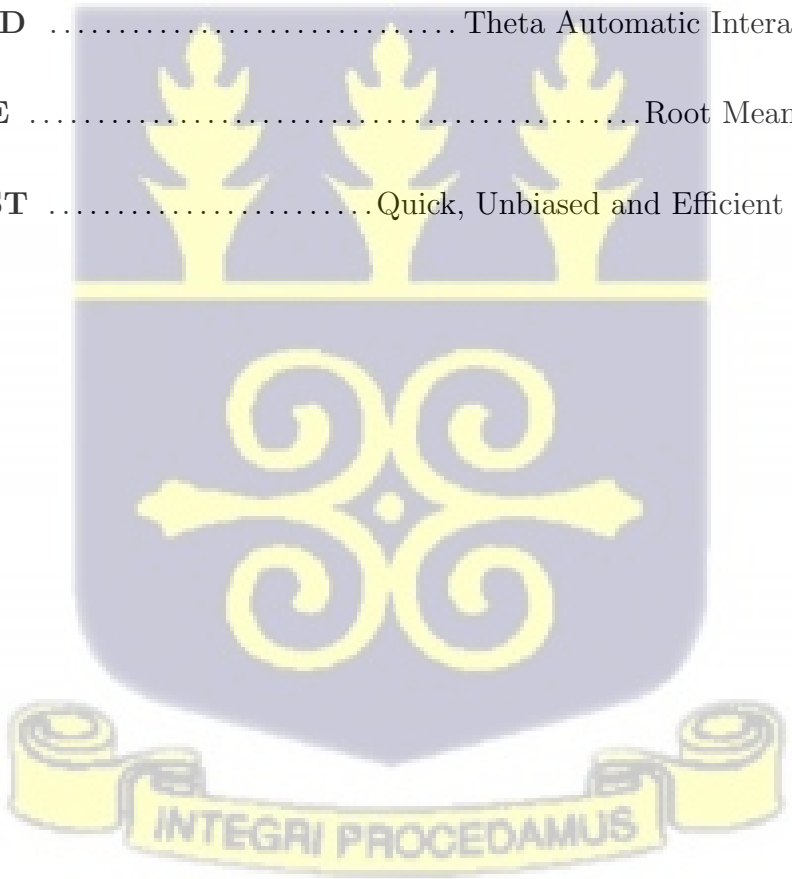
OG ..... Oil and Gas

PCC ..... Probability of Correct Classification

THAID ..... Theta Automatic Interaction Detection

RMSE ..... Root Mean Squared Error

QUEST ..... Quick, Unbiased and Efficient Statistical Tree



University of Ghana http://ugspace.ug.edu.gh

List of Tables

4.1 Summary Statistics of Insurance Claims Amount 45
4.2 Summary Statistics of Claims Amount: Insurance Cover 46
4.3 Summary Statistics of Claims Amount: Type of Entity 46
4.4 Description of predictive (explanatory) variables of Nigerian Dataset 46
4.5 Grid Search Result 49
4.6 Variable Importance 49
4.7 Final Results of the fitted CART tree (NG Data) 50
4.8 Optimum no. of mtry 52
4.9 Variable Importance 53
4.10 Relative Influence 55
4.11 Summary Statistics of Insurance Claims Amount (GH Data) . . . 58
4.12 Description of predictive (explanatory) variables of Ghanaian Dataset 59
4.13 Grid Search Result 61
4.14 Variable Importance 61
4.15 Final Results of the fitted CART tree GH Data 62
4.16 Optimum no. of mtry 63
4.17 Variable Importance 64
4.18 Relative Influence of the Variables 65
4.19 Performance Measures NG Data 69
4.20 Performance Measures GH Data 70

University of Ghana http://ugspace.ug.edu.gh

List of Figures

2.1 Partitions in a Decision Tree 12

4.1 Histogram of the Claims amount 45
4.2 Regression tree plot 47
4.3 The Optimal number of splits 48
4.4 Plot of the Complexity Parameter against the RMSE 49
4.5 Plot of Variable Importance (CART Model) 50
4.6 Optimum no. of Trees vs OOB Error 51
4.7 Randomly Selected Predictors 52
4.8 Variable Importance Plot 53
4.9 Selecting the best number of Trees 54
4.10 Relative Influence 55
4.11 Partial Dependence plot Variable Insurance Cover 56
4.12 Partial Dependence plot Insurance Cover 56
4.13 Partial Dependence plots Variable Year 57
4.14 Partial Dependence plot Insurance Type of Entity 57
4.15 Histogram of the Claims amount(GH Data) 59
4.16 Plot of the Final CART Model GH Data 60
4.17 The Optimal number of splits 60
4.18 Plot of the Complexity Parameter against the RMSE 61
4.19 Plot of Variable importance for the CART Model 62
4.20 Optimum no. of Trees vs OOB Error GH Data 63
4.21 Randomly Selected Predictors 64
4.22 Variable Importance Plot 64

University of Ghana http://ugspace.ug.edu.gh

4.23 Plot of the Optimum Number of Iterations 65
4.24 Relative Influence of the Variables 66
4.25 Partial Dependence Plot of the Variable Insurance Type 66
4.26 Partial dependence plot of the Variable Usage Type 67
4.27 Partial Dependence Plot of the Variable Purpose 67
4.28 Partial Dependence Plot of the Variable Purpose 67
4.29 Partial Dependence Plot of the Variable Vehicle Make 68
4.30 Partial Dependence Plot of the Variable Category 68
4.31 Partial Dependence Plot of the Variable Location 69

University of Ghana http://ugspace.ug.edu.gh

Chapter 1

Introduction

This study aims to conduct a detailed investigation into the viability of modeling of Non-Life insurance claims. The study aims to assess the predictive performance of Selected Machine Learning (ML) Algorithms to forecast Non-Life Insurance claims in the Nigerian and Ghanaian Insurance Market. Specifically, the selected ML algorithms are the popular Tree-Based models widely used to analyze insurance and Big data problems. They are the Classification and Regression Tree (CART), Random Forest (RF), and Gradient Boosting Models. Additionally, the study will carry out a comparative analysis of the selected Tree-Based models. It is important to briefly explain the key terminologies that would be used for the study. Hence, Non-Life Insurance, Claims, and Tree-Based analysis are introduced in the following paragraphs.

First, the non-life insurance mechanism entails providing an equal means of transferring the likelihood of a contingent or unpredictable liability in return for a monetary payment (known as a premium). As a result, the non-life insurance industry as a whole has a growing interest because it is expected to handle a vast number of conditions with a wide range of risks (Omari, Nyambura, & Mwangi, 2018). Therefore, the Non-Life insurance policies are, designed to provide coverage against uncertain events such as legal liability, accident, personal injury prevention, etc. In doing so, the insurer should be careful with the help of actuaries so that a fair premium is set and at the same time maintain the insurer's solvency and reduce the insurer's financial loss.

David and Jemna (2015) posit that insurance providers' main aim is to assess an

University of Ghana http://ugspace.ug.edu.gh

acceptable insurance amount or premium for an insured to meet a particular risk. Multiplying the conditional probability of claim frequency with the estimated expense or amount of claims is a well-known method of calculating the premium. Consequently, predicting claim amount, also known as claim severity or claim size, is a critical step in non-life insurance pricing. This forms the basis of this research and it will be discussed in detail in Chapter two. Second, According to Krishman (2010), a claim is a demand on an insurer to meet the portion of the commitment made when the insured made the contract. Furthermore, if insurers can overcome five main problems, such as having greater leverage over the claim process, considering their policyholders, selecting the best claims model for their business, establishing a mutually advantageous partnership with other service providers, and gaining a knowledge advantage, such an arrangement can become healthy. Hence the focus of this study would be on selecting the best claims model for the insurer's business. Third, for analyzing complex data structures, the ML algorithms approaches have become one of the most popular, intuitive, and efficient data analytic techniques available. The ML models offer a natural basis for forming individual subgroups for risk classification and regression models due to their many advantages. A decision tree model, according to Quan and Valdez (2018), is a data mining algorithm that can be broadly defined as repeatedly partitioning the regions of explanatory variables and thus constructing a tree-based model for predicting the outcome. Examples of the Decision Tree models are the ID3 (Iterative Dichotomiser 3), C4.5 (successor of ID3), Classification and Regression Tree (CART), Random Forests, Gradient boosting, etc. In chapters two and three of the study, the various concepts and theories underlying the foregoing terms will be explained in detail.

University of Ghana http://ugspace.ug.edu.gh

1.1 Background of the Study

The insurance industry has traditionally been an expanding sector. It plays a vital role in promoting a country's economic well-being as well as that of individuals and businesses. The policyholder makes a periodic payment of some certain amount called premium in exchange for financial protection against the possibility of losses in the future. When this loss occurs, claims are filed by the policyholder directly with the insurance company or by brokers or agents of the policyholder. Insurance claims service is usually intended to restore normalcy to the afflicted and insured individual or business. And, for the insurer to create a favorable environment for the efficient provision of this non-life insurance service, it needs the right premium for the customer at the beginning of the insurance contract. This is necessary for an insurance practice. Therefore, it is critical to have an accurate and reliable estimate of the number of claim occurrences and the size of those claims.

Consequently, estimating and predicting these insurance claim size is one of the major challenges insurance companies face. Actuaries are tasked with the prediction of insurance claims and identifying what risk factors are responsible for the determination and prediction of insurance claims using various statistical methods (Edward & Valdez, 2012). This is important for the insurance company because they are extremely concerned about forecasting the future claim size. Accurate forecasting increases the likelihood of reducing the insurance company's financial loss and also allows them to set adequate reserves (Burri, Burri, Bojja, & Buruga, 2019). For this reason, insurers use very complicated methodologies. A decision tree, a random forest, a binary logistic regression, and a support vector are some of the main models which actuaries use to predict insurance claims. This research plan to focus its attention on the Decision Tree in the prediction and estimation of the claim amount of the non-life insurance.

University of Ghana http://ugspace.ug.edu.gh

The Traditional Decision tree dates back to the 1960s. Specifically, Morgan and Sonquist (1963) later developed the regression tree called Automatic Interaction Detection (AID). The Decision Tree is a data mining method that works by frequently splitting the domains of predictors and building a tree-based model to predict the result. Before the decision tree, the GLM was the popular method for predicting insurance claims. Prior researchers made significant efforts in trying to develop insurance claims modeling using the popular GLM and other statistical models like the Spatial Model, Tweedie's Compound Model, Hierarchical Models, etc. The main focus here is getting the best model capable of predicting the insurance claims amount so that the insurer can produce or determine fair and accurate tariffs or premiums and also ensure the financial stability of the insurer (Smyth & Jorgensen, 2002).

Omerašević and Selimović (2020) outlined that the GLM is a common statistical model in the actuarial mathematics of non-life insurance. They also identified British Statisticians Nelder and Wadderburn (1972), with developing the GLMs in both actuarial science and statistics. They illustrated that GLM is an expansion of conventional linear models, in which the dependent response variable's probability distribution is a part of the family of exponential distributions (normal, gamma, Poisson, etc.) and the dependent variable's expectation is estimated using a linear model based on a nonlinear relation function.

The GLM has some benefits over other techniques for modeling insurance claims that have been used historically. One of them is the general statistical framework, which includes techniques for calculating standard error, the goodness of fit, and confidence intervals, among other things. Furthermore, basic statistical tools for the GLM, such as SAS, R, SPSS, and others, allow data analysis for making forecasts and determining premium rates relatively simple. But recent research shows that Decision trees are now gaining more popularity than the GLM. The

University of Ghana http://ugspace.ug.edu.gh

decision tree due to its numerous advantages, is gaining more popularity as a data mining algorithm for classification, prediction, and analyzing actuarial data and does not require distribution assumptions. On this basis, the decision tree model has several advantages over the GLMs.

Firstly, since a decision tree model is nonparametric, no distribution assumptions are required. Decision tree models, unlike standard statistical approaches, do not require any data from probability distributions concerning the outcome. Secondly, it is a powerful algorithm that can deal with lost data. It's not unusual for any details to be lost or unrecorded in actual datasets. Third, it can detect non-linear effects and potential correlations within the explanatory variables, in addition to handling missing details. Traditional linear models usually absorb only linear effects, necessitating further testing to find non-linearity and interactions. Fourth, by measuring the relative value of the explanatory variables, it can be called a variable selection method. Variable selection is common in actuarial science for applications such as risk classification and risk variable selection. Lastly, decision trees are simple to interpret using a simulation of the tree structure in the plot, particularly with smaller trees. These benefits are extremely relevant for actuarial and insurance data (Tan, Steinbach, & Kumar, 2006; Quan & Valdez, 2018; Omerašević & Selimović, 2020).

But on the other hand, Omerašević and Selimović, (2020) posit that the decision trees do not have a variety of analytical checks and statistical measurements, but the statistical tools like GLM have. Furthermore, if the dependent and the independent variable have a linear relationship, predictive approaches outperform the decision tree. However, when used in the data analysis process for further use in a statistical model such as GLM, the decision trees, on the other hand, do not have any special restrictions or conditions.

Quite a number of researchers have attempted to investigate insurance claims of different insurance classes of business, but the scarcity of data and the difficulty

University of Ghana http://ugspace.ug.edu.gh

involved in obtaining it from the insurance companies in Nigeria, is posing a challenge in modeling individual non-life insurance claims. The only data readily available to the public or researchers is the basic detailed statistics performed annually by organizations such as the Nigerian Insurance Association (NIA). Using these statistics, a few studies like Adebayo and Sunday (2017) examined the insurance claims of the Oil and Gas Insurance Class of Business using Extreme Value Theory. Dallah (2010) analyses the Armed Robbery Insurance claims in Nigeria using an expert model mining system. Chukwudum (2018) demonstrates how extreme value theory can be applied in modeling the probabilistic behavior of the frequency and severity of large motor claims from the Nigerian insurance sector using the Negative Binomial-Generalized Pareto distribution. Oyewole, Adesina, and Adekola (2018) also investigated the Nigeria Marine and Aviation Insurance class of Business. This study used the Extreme Value theory to estimate the minimum expected claims for marine and aviation insurance businesses based on historical claims data. None of these studies adopted the Tree-Based Models in the analysis of insurance claims.

It is worthy to note that some market conditions such as broker participation levels, caps on liability, the Nigeria Insurance Act, etc. may affect the nature of the distribution of the insurance claims and may have a strange relationship with the covariates that make one modeling method more accurate than the other. An example of market conditions is evident in Omerašević and Selimović (2020) where the researchers outlined that there some changes in laws related to the liberalization of the motor third-party liability insurance industry, and these changes eventually resulted in the insurance market in Bosnia and Herzegovina contracting. The researchers also outlined that experiences of neighboring Croatia and Slovenia, two years after price liberalization, the demand for motor third-party liability insurance fell by 30%, resulting in major losses for the insurance sector in these countries. Unlike other insurance markets, the Nigerian insurance market has been characterized by some major challenges. This includes low

University of Ghana http://ugspace.ug.edu.gh

penetration levels, lack of customer trust, low implementation of compulsory insurance, and lack of professionals that are adequately skilled in this space (PWC Report, 2015).

This study, therefore, aims to determine the appropriate models and approach for the prediction of claims amount of the Nigerian Insurance market and then compare it to the Ghanaian dataset to test the performance of the selected models.

1.2 Problem Statement

Generally, insurers require the use of claims data compiled from previous years for rate-making purposes, setting of reserves, arranging for reinsurance, and other reasons. They achieve this through modeling the gathered claims data. Actuaries use complex methods due to the importance involved in predicting and forecasting insurance claims because accurate forecasting greatly reduces the probability of financial loss on the insurer (Burri, et al., 2019). In other words, precise forecasting assists insurers to adequately cover anticipated risks, costs, contingencies, and ultimately determine insurance premium. Several previous researchers modeled non-life insurance claims (Ohlsson, & Johansson, 2010; David, & Jemna, 2015; Shi, Feng, & Ivantsova, 2015; Weerasinghe, & Wijegunasekara, 2016; Wüthrich, 2017; Omari et al., 2018). But classifying a particular situation as belonging to a certain class is of great significance. Hence the problem of accurate prediction in insurance claims is expressed as a classification problem (Weerasinghe, & Wijegunasekara, 2016). For insurers to be able to resolve claims that could emerge from existing portfolios of policies at any stage in the future, past and current information on claims history must be accurately modeled; these can be used to forecast anticipated future claims experience and set appropriate reserves. When modeling claims data, non-life insurance firms often face two challenges: choosing suitable statistical distributions for claims data and determining how well

University of Ghana http://ugspace.ug.edu.gh

the chosen statistical distributions fit the claims data. Therefore, insurers would have to rely upon the best approach to price their insurance contract adequately. To do this, insurers need to model the frequency and severity of claims based on the popular assumption that the claim frequency and severity are independent. Therefore, traditional models, such as count regression analysis, allow for the detection of risk factors as well as the calculation of the predicted occurrence of claims based on those risk factors.

In addition, the commonly used distributions and approaches for modeling claim frequency and severity cannot accurately represent real claim data distributions, therefore, this necessitates some changes to standard distributions. Most insurers depend on developed-market models that are typically tailored to simulate their claim frequency and severity distributions. Since all markets have unique characteristics, individual models from different markets cannot be extended to accurately model and predict the claims data of all insurance providers.

Therefore, the interest of the researcher in this study is to use the Tree-Based Models as an empirical approach to predict the amount of non-life insurance claims in the Nigerian and Ghana Insurance market.

1.3 Objectives of Study

The main objective of this study is to investigate the characteristics of the non-life insurance data using Machine Learning (ML) models. The specific objectives are:

- To predict insurance claims amount using the Classification & Regression Tree (CART), Random Forest (RF), and Gradient Boosted (GB) regression models.
- To carry out a comparative study between the selected ML models for the

University of Ghana http://ugspace.ug.edu.gh

prediction of insurance claims amount.

- To select a suitable statistical model which accurately predicts and describes the claim amount of the non-life insurance data.
- To identify the risk factors that are significant in determining claims amount.

1.4 Research Questions

The research aims to answer the following questions to model and predict the claim amount of non-life insurance using the best approach. The research questions are;

- Does ML models effectively handle data characteristics?
- What approach accurately predicts the claim amount of the non-life insurance?
- What approach best describes the claim amount of the non-life insurance?

1.5 Scope and Limitations

This study aims at determining the model with the best predictive accuracy between the CART, RF, and GBM. The data that would be considered for this research is the insurance claims amount Covering all insurance companies in Nigeria from 2011 to 2018. The Ghanaian data on the other hand consist of motor vehicle insurance for the period 2013 to 2016.

University of Ghana http://ugspace.ug.edu.gh

1.6 Significance of Study

To the best of our knowledge, most of the previous studies applied the probability distributions in modeling insurance claims in the Nigerian Insurance industry. Therefore, the results of this research will add to the scarce amount of literature on the application of Tree-Based ML models in the Nigerian insurance market. This is important for the management of the insurance company in a highly competitive economic environment to set a fair insurance premium and to know which of the risk factors can be incorporated in the premium. Secondly, it will assist the insurance company to predict the amount of claims given the risk factors. This would allow insurers to analyze their claims processes based on historical data and make informed decisions to enhance efficiency. In a higher context, the application of the considered models to non-life insurance claims provides a practical framework of the Tree-Based Analysis in Actuarial Science and Insurance. Furthermore, the study will complement the research effort of the Nigerian actuaries whom would like expand their knowledge on ML algorithms to precisely and accurately estimate insurance claims.

1.7 Organization of Study

The study will be organized into five chapters. Chapter one provides the background of the research. Chapter two gives the literature review and model assessment of other researchers in the field. The methodology is discussed in chapter three. Chapter four will comprise of analysis and presentation of results including interpretation of results. Lastly, chapter five discusses the research findings, conclusions, and recommendations.

University of Ghana http://ugspace.ug.edu.gh

Chapter 2

Literature Review

This chapter includes the review of relevant literature on selected Tree-Based ML Models. The chapter will be divided into four sections, each of which will examine a different aspect of the research. Section one provides an overview the Decision Tree and of its other methods, as acknowledged in Actuarial Science and Insurance. The background and application of the Models are explored in sections two and three respectively. Lastly, the modeling of claims that are peculiar to the Nigerian insurance market were discussed.

2.1 Decision Tree

A theoretical structure of decisions and underlying effects is referred to as a decision tree. It is also a predictive model that may be used to define both classifiers and regression models. The decision-maker uses decision trees to determine out which plan is most likely to succeed (Maimon & Rokach, 2014). Decision trees are straightforward but effective methods for forecasting and describing the link between various measures about an object and its desired value. Decision trees, which originated in management, logic, and statistics, are now very successful tools in other disciplines such as information extraction, text mining, pattern recognition and machine learning, in addition to data mining (Maimon & Rokach, 2008). Morgan and Sonquist (1963) created the first regression tree system, which they called Automatic Interaction Detection (AID).

Since Breiman et al. (1984) developed the Classification and Regression Tree

University of Ghana http://ugspace.ug.edu.gh

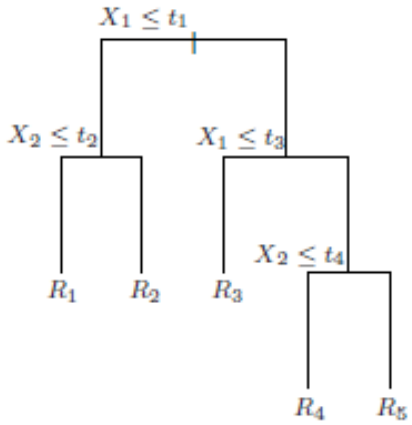


Figure 2.1: Partitions in a Decision Tree

(CART) approach, tree-based analysis has gained popularity in research and applications in a variety of disciplines. The CART algorithm separates the explanatory variable space into separate mutually exclusive sections, resulting in a tiered pyramid of branches resembling a tree assembly, each branch being referred to as a node. Each of the decision tree's bottom nodes, known as terminal nodes, has a distinct path for data to reach the region. Following the construction of the decision tree, pathways may be used to find the area or region to which a fresh collection of explanatory variables will fit. The decision tree method may be split into two types based on the aim of predictive modeling: regression and classification trees.

Figure 1 above is an example of a decision tree where the complete datasets sits at the top of the tree. But all the observations meeting the criteria at each split are assigned to the left branch, while all others are assigned to the right branch. In the above exmaple terminal nodes are represented by the regions $R_1, R_2, ...R_5$.

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

2.2 Theoretical Background of the Tree-Based Models

Tree-based machine learning models have gained popularity as a strong technique for decision-making in various fields since Breiman et al. (1984) developed the Classification and Regression Tree (CART) algorithm. The CART algorithm divides the explanatory variable space into many mutually exclusive areas, resulting in a layered tree-like hierarchy of branches. A node is a point where two lines or branches meet. Each of the decision tree's bottom nodes, known as terminal nodes, has its path for apparent data to enter. Paths can be used to find the area or terminal node to which a new collection of explanatory variables will fit after the decision tree has been created.

Generally, a regression tree can be mathematically represented as follows: given a vector of covariates $x^* \in R^p$, its dependent variable, $y^* \in R$, is predicted as:

$$f(x^*) = \sum_{q=1}^{Q} I(x^* \in g_q)\mu_q \tag{2.1}$$

where $g_q \subseteq R^p$, $\mu_q \in R$ for all $q \in \{1, ..., S\}$. $\{g_q\}_{q=1}^{Q}$ defines a partition of the covariate space, that is $g_q \cap g_r = \emptyset$ for $s \neq t$ and $\cup_{q=1}^{Q} g_q$ is the covariate space. μ_q Is the predictor of dependent variable realizations, whose covariate vectors are in g_q. The nodes $\{\mu_q g_q\}_{q=1}^{Q}$ are selected using a recursive optimization process that can be diagrammed as a tree, where S are terminal nodes (i.e. the leaves). The recursive optimization for selecting S is based on the training set, $\{y_i x_i\}_{i=1}^{n} = \{y, X\}$ where $y \in R^n$ and $X \in R^{n \times p}$ is the design matrix, and it is assumed that x^* and y^* were drawn from the same distribution as x_i and y_i Each step in the tree's recursive optimization process is a model selection of linear

University of Ghana http://ugspace.ug.edu.gh

models problem, where a threshold of one of the available covariates is selected for minimizing a loss function $\text{Loss}(\cdot,\cdot) : R^2 \to R$

2.2.1 The Concept of Tree-Based Model Algorithms

According to Quan (2019), the concept of the tree-based approach assumes that the dataset comprising a vector of k explanatory (or independent) variables which is given by $\mathbf{X} = (x_{i1}, x_{i2}, ..., x_{ik})$ sampled from the space $X = X_1 \times \cdots \times X_k$ and a response (or dependent) variable, y_i, sampled from the space Y, for the observations of each N. For $i = 1, ..., N$ the ideal representation of this data is $(\mathbf{X_i}, y_i)$. Under this concept, several models have been developed over the years. Below is a list of

- Classification and Regression Tree (CART) (Breiman et al., 1984)
- Chi-Squared Automatic Interaction Detector (CHAID) (Kass, 1980)
- C4.5 an extension of the ID3 algorithm (Quinlan, 1986)
- Fast and Accurate Classification Tree (FACT) (Loh & Vanichsetakul, 1988)
- Quick, Unbiased and Efficient Statistical Tree (QUEST) (Loh Shih, 1997)
- Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE) (Kim & Loh, 2001)
- Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) (Loh, 2009)

The Ensemble, and other Methods

- Random Forest (Breiman, 2001)
- Bagging (Breiman, 1996)

University of Ghana http://ugspace.ug.edu.gh

- Boosting (Freund & Schapire, 1997)

However, the most extensively applied Univariate Regression Tress are discussed in the subsequent subsections: Classification and Regression Tress, Bagging and Random Forests, and Gradient Boosting.

2.2.2 Classification and Regression Tree (CART)

A Classification and Regression Tree (CART) is a machine learning predictive model. It illustrates how the values of a target variable can be predicted using other values. It's a decision tree with each split separated into a predictor variable and a prediction for the target variable at the end of each node. It uses the same greedy search strategy as Automatic Interaction Detection (AID) and Theta Automatic Interaction Detection (THAID) but improves a few advanced functionalities (Breiman et al., 1984; Loh, 2014). It develops a big tree and afterward prunes it to the size that takes the minimum cross-validation assessment of error, rather than applying termination rules. The type of application of the decision tree approach are in two forms. First, if a variable is categorical, the decision tree is applied for classification purposes. On the other hand, if a variable is continuous, it is applied for regression. Tober (2020) divides CART into two parts: one for growing and the other for pruning. Given a node m, a candidate for a split or separation condition s, and a variable X on which the split is performed. CART then separates m into a right node, given by $m_R = \{X : X > s\}$ and a left node, $m_L = \{X : X \leq s\}$ so that we can have $m = m_R \cup m_L$ This process is then repeated until a stopping requirement is met, such as the tree's maximum depth. In general, s and X are chosen so that, given the two daughter nodes, some loss function, $L(\cdot, \cdot)$ is minimized (Freidman et al., 1984).

The CART also involves surrogate splitting, which entails searching a surrogate variable which adequately approximates the initial split, and can also be used by

University of Ghana http://ugspace.ug.edu.gh

regression trees to manage omitted values in explanatory (independent) variables. Surrogate splitting, in other words, seeks out a substitute explanatory variable that closely resembles the missing-value explanatory variable (Quan, 2019). As a result, the datasets with omitted values are allocated based on the surrogate variable's split rather than the original splitting variable's split. By monitoring the decrease in the error function, the surrogate splitting mechanism offers varied importance to each node. Klusowski (2019) proposes that for one to use CART, one must first decide on a loss function. The loss function would be explained the methodology section of this thesis.

2.2.3 Bagging and Random Forests

Bagging, developed by Breiman (1996), is based on the observation that a fraction of the outcome error in a particular regression tree can be attributed to the training dataset's precise selection. As a result, the variance element of the outcome error is minimized after numerous identical datasets are constructed using resampling with replacement (bootstrapping) and the created regression trees remain without pruning and averaging (Prasad, Iverson, & Liaw, 2006). Bagging, in other words, employs an ensemble of relatively robust, unpruned CART trees built via bootstrap sample from the initial training dataset. The unpruned CART can be given as $\{Q(X;\Theta_b), b = 1, ..., B\}$. B is defined as the aggregate sum of bootstrap samples and the framework for the dependent variable would be equal to the mean of all the regression trees in bagging.

According to an article by Breiman (2001): "a random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1, ...\}$ where the Θ_k are independent and identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$."

Quan (2019) further posits that the idea of Random Forest was to minimize the

University of Ghana http://ugspace.ug.edu.gh

link bwetween the Classification and Regression Trees, so that an optimum number of split is selected at random from a set of independent variables at every node of the regression tree. Another option is to choose the training set from a random set of weights applied to the training set's examples. Cross-validation can be used to determine the best size for the randomly chosen subset of Independent variables. In practice, the average forecast of many Classification and Regression Trees should have reduced variation than an individual Classification and Regression Tree. Broader randomly chosen sets of independent variables tends to increase individual trees' predictive ability, but they also enhance correlation across trees, eliminating any advantage from averaging many estimates. The bootstrap resampling of the training dataset for each tree enhances the heterogeneity amongst the trees as well. Random forest accuracy is determined by the robustness of each tree and the degree to which they are interdependent.

2.2.4 Gradient Boosted Regression Trees

Gradient-boosted decision trees developed by Friedman (2001) are a machine-learning approach to optimize the explanatory power of a model via continuous learning stages. Each stage of the decision tree entails altering the coefficient values, weights, or bias values used to estimate the optimal model of each of the response variables to decrease the prediction error (deviation measurement between the expected values anticipated and observed). In other terms, regression trees are built using the originally created regression trees information so that every regression tree subsequently concentrates upon learning about the residues from prior trees (Quan & Valdez, 2018). The gradient is the stepwise adjustment performed at every stage of the process; boosting is a way to accelerate predictive accuracy improvements to a reasonably high level. According to Breiman (2001), the result of the tree would be a collection of J trees given as $T(X;\Theta_J)$, for j = 1, . . . , J.

University of Ghana http://ugspace.ug.edu.gh

2.3 Empirical Application of the Tree-Based Machine Learning Models

There is a large range of interdisciplinary applications in tree-based ML models. The approach has been applied in various disciplines such as Mathematical Sciences, Engineering, Computer Sciences, Biological Sciences, Geographical Sciences, Econometrics, Information Technology, Business & Market Research, etc.

In Marketing and Business research, Durica, Frnda, and Svabova (2019) utilized the tree-based models (CART and CHAID) to predict business failure for polish companies. The study investigated the financial records and accounts of companies operating in Poland for 2016 and 2017 based on the Amadeus database. The modeled data comprises 37 predicted financial indicators. Interestingly, they find that the approach obtained a predictive accuracy of 98 percent for companies. Non-prosperous enterprises' classification success is more than 83 percent. The Decision-Based Prediction Models are especially effective in anticipating Polish enterprises' financial problems, they can also be utilized in other countries.

In another article by Obiedat (2020), the researcher focuses on the Decision Tree (DT) algorithms to forecast the efficiency campaigns using data from a Portuguese bank's marketing section. CART, C4.5, Random Tree, and RepTree were the algorithms used. The results reveal that Simple CART is the optimal method for predicting the success of campaigns. In the study, however, a filter-based functional selection was utilized to increase classification performance.

In a combined investigation in the field of Geographical, and Environmental Sciences, Anderson, Guikema, Zaitchik, and Pan (2014) employed both regression and the Tree-based models to anticipate, predict and estimate population densities vital for policy and health planning based on Peruvian population data between the periods 1993 to 2007. The results demonstrated that non-parametric

University of Ghana http://ugspace.ug.edu.gh

tree models provided better accurate predictors than traditional regression methods.

In the field of Electrical Engineering, Jamehbozorg, and Shahrtash (2010) introduced a new approach for detection of fault within a singled-circuit transmission system. Furthermore, the scholars were able to detect the fault type using a Tree-based analysis and the results show high accuracy in terms of classification of faults.

An Accident Analysis and Prevention Research conducted by Yan, Richards, and Su (2010) employed the Hierarchical tree-based model. The goal of the research was to forecast train accident crash predictions and assessment at an inactive major railway crossings based on the Federal Railroad Administration (FRA) data in the United States for the periods 1980 to 2006. Applying the non-parametric technique, the researchers were able to forecast the frequency of train vehicle accidents for inactive railway crossings. The results reveal the solution upon how to increase safety at inactive railway crossing.

The above citations are just a few examples of the applications of the Tree-based models in other disciplines. For more examples, see Shannon et al. (2001), Piboolnurak et al. (2005), Jo et al. (2007), Esposito et al. (2013), Singh et al. (2016), Kadiyala, and Kumar (2018), Ter Avest et al. (2019), Kershaw (2019), Ter Avest (2019), Kaparthi, and Bumblauskas (2020), Zounemat-Kermani et al. (2020), Neumann and Klatt (2021). All these and many more can be found in recent literature.

INTEGRI PROCEDAMUS

On the other hand, the field of Actuarial Science and Insurance witnessed the Generalized Linear Models (GLM) as the most popular statistical models applied. The Tree-Based ML Models are also gaining more popularity in recent times. Most importantly the reason for its popularity can be attributed to its advantages.

University of Ghana http://ugspace.ug.edu.gh

There is a need to appreciate the general merits of the Tree-Based ML models over the standard statistical and other data mining methods. As pointed out by Omerašević, and Selimović, (2020), and Quan (2019), the Tree-based ML models has the below merits summarized as thus:

1. Decision trees are fast and easily constructed to provide straightforward and comprehensible outcomes.

2. The graphic display of data as a tree helps to comprehend the causative factors of the response variable predictors and it can also handle variables with missing data values.

3. The autocorrelation of the variables does not affect the decision tree performance since the majority of decision tree algorithms are highly efficient and frequently have enormous data.

4. Decision tree templates are especially suited for decision making when the predictor's relationship with the response variable is not previously informed and they produce very precise outcomes for prediction.

Having stated the above general advantages, the application and importance of the Tree-Based ML approaches in actuarial science and insurance literature are beyond the predictive accuracy of the models, but rather it can also be combined with traditional models to improve the standard general statistical framework, which establish the increased performance of such models, and also overcome their shortcomings, eliminate categories with insufficient data, and detect for hidden patterns in the data to obtain very vital information (Omerašević, & Selimović, 2020). In addition to this, the traditional statistical software like the SPSS, EVIEWS, R, etc. for the models makes the analysis of data easier.

Consequently, since the presentation of the traditional GLM, a wealth of excellent work has been published that has enabled several authors and researchers to

University of Ghana http://ugspace.ug.edu.gh

identify, formulate or strengthen the hypothesis that has permitted the use of these models in non-life insurance especially in the determination of insurance premium and/or pricing of insurance contracts.

We now discuss some of the applications of the Tree-Based ML models in actuarial science and insurance.

For instance, Guo (2003) was one of those researchers who presented two significant insurance data mining algorithms: decision tree analysis and cluster discovery techniques. The researcher in this study outlined the relevance of Tree-Based analysis to actuarial practice and insurance. Using a simple dataset regarding a list of drivers, the researcher described how to evaluate the number of claims in the data for underwriting. The researchers further added that the evaluation would enable the actuary to assess the potential profit of prospects from the list of drivers/policyholders.

Using large claims datasets, Derrig and Francis (2006) analyzed the relative performance in the prediction of important claim variables besides excessive and dishonest practices and expectations for favorable results from the investigation using the various decision tree-based analyses such as CART, TreeNET, Random Forest, S-PLUS, and Insightful Miner Tree Procedures. The data (roughly about 500,000) was utilized to analyze the car injury claims reported in detail is a dataset from the auto insurance office of Massachusetts from 1995 to 1997. Results of the study demonstrate that tree-based approaches have some predictive power or lift the variables with substantial variations between methods. However, all of the modeling results are comparable to the logistic regression identical to Viaene et al. (2002).

In the field of expert auto insurances fraud detection, Viaene et al. (2002) assess different state-of-the-art binary classification approaches. The researchers evaluated the predictive accuracy of the logistic regression, the C4.5, the k-nearest

University of Ghana http://ugspace.ug.edu.gh

neighbor, the Bayesians learning multi-layer perceptron neural network, and the least-squares vector machine. Their results suggest that the incorporation of non-flag predictors considerably increases predictive performance, regardless of target encoding system and algorithm type. Furthermore, In addition, for all scenarios assessed, the mean Probability of Correct Classification (PCC) and mean Area under the Curve Receiver Operating Characteristics (AUCROC) performance differential amongst numerous algorithm operations is quite low.

Olbricht (2012) studied the life table construction applied the Tree-based method using a Swiss Re dataset as an example. The dataset consists of an age range spanning from 0 to 65 with 1,463,964 records for four years with two years for the training dataset and two years for testing. The researcher states that these tree-based approaches contain positive attributes for the identification of primary risk variables via the intrinsic data structure.

Guelman (2012) investigated the auto insurance loss amount with the aid of the gradient boost tree-based approach. The analytical data for the study were collected from a huge database by a major Canadian insurance company which comprises information about policies and claims at the level of the vehicle for the period Jan 2006 to Dec 2009 with a total of 14,984 claims. The researcher divided the dataset into two; 70 percent for the training dataset and 30 percent for the testing dataset. As discussed earlier, based on the independence assumption, the researcher in this study analyzed the number (frequency) of claims and amount (severity) of claims separately. The researcher concluded that the gradient boost method gave better prediction accuracy than regular GLM.

Guelman, Guillén, and Perez-Marin (2014) conducted a study to select profitable policyholders using the probabilistic inference trees to personalize cross-sell advertisements of insurance products. This was carried out by building personalized treatment learning using a email marketing advertisement launched by a major Canadian insurer between June 2012 and May 2013 to cross-sell home insurance

University of Ghana http://ugspace.ug.edu.gh

to some selected current automobile insurance customers.

Deprez et al. (2017) harnessed the tree-based analysis to estimate the conditional probabilities in a Poisson model and demonstrate how the machine learning tree-based analysis can be used to investigate the accuracy of mortality approaches to appreciate several reasons of death. The research was based on Swiss mortality data. The researchers concluded that the boosted regression tree can be utilized to back-test parametric models. The researchers added that the tree-based approaches allow for the detection of weaknesses of the models based on real data. Furthermore, regression tree boosting can be used to optimize the fitness of such modeling techniques for feature variables that aren't incorporated by the algorithms.

Lee and Lin (2018) proposed Delta Boosting (DB) as an alternate boosting method to the Gradient Boosting (GB), demonstrating that it is optimal for a wide range of loss functions. The study was based on a dataset from a Canadian Insurance company which comprises claims and policy information of vehicles for collision coverage for accident years 2001 through 2005. The researchers discovered that, like the GB, the DB provides users with a high level of transparency, and that the marginal impact of variables may be examined using statistical significance diagrams and partial dependence graphs. They further added that the lifts, losses, and double lifts on the sample size can be used to evaluate the overall performance of the model.

Dewi, Murfi, and Abdullah (2019) carried out an analysis accuracy of random forest to predict the amount of car insurance claim severity. To demonstrate the model's high predictive performance, the scholars utilized only 1/3 of the overall variables or features. The scholars used the Allstate insurance data having an observation of 188,318. The data is readily available at kaggle (https://www.kaggle.com/c/allstateclaims-severity). The scholars used the training data to create a method which has the ability to forecast the severity of an

University of Ghana http://ugspace.ug.edu.gh

insurance claims when a car accident occurs. The scholars further, estimate the predictive performance of the methods created in predicting insurance claims from the testing dataset. The scholars concluded that RF model is applicable to prediction of insurance claims which is a case of regression problem in the light of ML Algorithms with a predictive accuracy of 99%.

Omerašević and Selimović (2020) in a study to analyze the effect of risk classification on life insurance pricing using the Tree-based approach investigates the insurance policies of passenger vehicles for the years 2015 to 2020. The researchers propose a standard risk approach based on the application of GLMs with Poisson for the prediction of the number of claims and Gamma for the prediction of claims amount. According to their findings, the results show that GLMs do not always offer credible parameter estimates for categorical predictors with multi-level factors (MLF).

Quan et al. (2020) proposed the tree-based approach with such a hybrid nature adopting a two-step method as a different approach to the classic GLMs. To design the probability model for frequency, the scholars outline that the first step is to create a classification tree. In the second stage, the scholars utilize the elastic net regression models to create the severity distribution model for every terminal node from the classification and regression tree. The scholars compare the Tweedie GLM with the hybrid tree-based approaches based on the LGPIF data spanning from 2006 to 2011. Their findings suggest that hybrid tree-based approaches offer greater precise forecasts while sustaining intuitive interpretation.

In recent research by Henckaerts et al. (2020) aimed at developing a full tariff plan (insurance pricing), the researchers adopted a process such that the price of insurance is built using the tree-based analysis based on the number of claims and amount of claims. In this article the loss functions in the algorithms have been utilized to properly incorporate the special characteristics of insurance data: high-unbalanced count data with extra zeros and fluctuating frequency exposure

University of Ghana http://ugspace.ug.edu.gh

mixed with limited but perhaps long-tailed data on the severity side. Using a motor third-party liability dataset from a Belgian insurance company for the period 1997, the researchers concluded that the Boosted Tree approach has higher predictive accuracy than the traditional GLM.

A more recent study can also be seen in Staudt and Wagner (2021) where the researchers employed the Random Forests for modeling claim severity based on the Collision in Car Insurance dataset for the period 2011 through 2015. The researchers followed the standard independence assumption between the frequency and severity models for computing non-life insurance premium, which actuaries normally rely on. Since the focus of their research was placed on the severity of claims, the first step was to develop two reference models a GLMs and GAMs, the second step was to tune two random forests models. The models were compared with each other utilizing the root mean squared error, relative error, and goodness-of-fit statistics, to assess their predictive performance. The scholars concluded that the use of a log-normal transformation of the severity does not improve random forests performance. They further added that the GAMs emerged as the best model with the highest predictive accuracy.

2.3.1 Insurance Claims Modeling in the Nigerian Insurance Market

Having reviewed some relevant literature concerning Insurance claims using the Tree-Based ML algorithms and other approaches applied to the Insurance market data of both the developed and developing economies, it would be of importance to have a look at the Nigerian Insurance market data since the current research is aimed at investigating insurance claims data in the Nigerian Insurance market which forms the basis and focus of this research. Furthermore, it would enable us to determine what has been done, find a gap in the research, and fill this

University of Ghana http://ugspace.ug.edu.gh

gap. However, an overview of the literature concerning the analysis of insurance claims in the Nigerian Insurance market reveals that not much research has been carried out in this area of Insurance. To the best of our knowledge, the following pieces of literature are the only available studies carried out concerning modeling insurance claims in the Nigerian Insurance market.

Dallah (2010) carried out a study to model the armed robbery claims in the Nigerian Insurance Industry. Using the claims data of Wapic Insurance Company, the researcher investigated only a single aspect of insurance portfolio for the period 2000 to 2004. In an effort to develop and accurately determine the price of insurance premium of the affected portfolio, the researcher adopted some standard statistical approaches such as the Q-Q plots, the P-P Plots, the Kolmogorov-Siminov Test, and the Three-Parameter Lognormal distribution to carry out the study. The researcher attempted to predict the extreme claims risks events in the portfolio analyzed. The researcher posit that the study's findings were extremely informative and also in accordance with earlier researches conducted in a similar field which presented the Lognormal distribution as one of the candidate model that can fit into the analysis of insurance data that has a right skewed distribution.

Adeleke and Ibiwoye (2011) investigated the claim amounts based on the all the Non-Life Insurance classes in Nigeria to find a suitable approach or methods to model those claims. The scholars carried out this research through fitting the probability distributions on the large claims data obtained from the Nigeria Insurer Association (NIA). The distributions employed by the researchers included the Exponential, Pareto, Gamma, Weibull, and the lognormal distributions. The researcher predicted the premium risk of all the Categories of insurance in the Nigerian Insurance Industry and concluded that certain categories of insurance claims are best modeled using alternative distributions. In their conclusion, the researchers stated that Gamma distribution is best for Commercial, Property, and Fire insurance policies, lognormal for Motor and Theft insurance policies,

University of Ghana http://ugspace.ug.edu.gh

and Weibull for Armed Robbery insurance policies.

In another article by Adebayo and Sunday (2017), the researchers analyzed the Oil and Gas Insurance claims category in the Nigerian Insurance market using the Extreme Value Theory (EVT) approach. This approach was considered to be appropriate for modeling such claims. The researchers carried out this study based on a sample of data for 5 insurance companies on Oil and Gas Insurance claims category from 2011 through 2012. Specifically, the models adopted were the Generalized Pareto Distribution. The Maximum Likelihood estimate was the statistical technique adopted to estimate the model parameters. Other Statistical tools included the Q-Q Plots and Grobb's test. Furthermore, the results of the study obtained were contrasted with the Gaussian and Historical approach and concluded that the Extreme VaR was the correct method to compute the VaR.

In an attempt by Oyewole, Adesina, and Adekola (2018) to quantify tail risk in finance and business, the authors in this research applied the Extreme Value theory to analyze extreme losses in the Nigerian Marine and Aviation insurance claims category. This study was based on the claims data of four insurance companies for a period of two years, 2011 to 2012. In this research, the three VaR estimation approaches were investigated and the empirical outcome demonstrates that extreme VaR is best for Historical and Gaussian approaches. This research provided an estimate of the least amount of claims which the insurance company would pay in a specified period based on the EVT approach. The researchers determined that the Extreme Value Theory technique of computing VaR surpasses other estimating approaches since it is acknowledged for its superior ability to predict the distribution's tail area. Other approximation approaches function well when dealing with high-profile data, however severe VaR is estimated regardless of the number of data sets.

The most recent study carried out concerning insurance claims in Nigeria was that of Chukwudum (2019). In this study, the scholar studied the Motor Insur-

University of Ghana http://ugspace.ug.edu.gh

ance claims categories about large claims. The scholar attempted to model the number of claims and the size of claims with their respective behavior under the probability distributions using the GLMs to describe how Extreme Value Theory can be employed for the analysis. The study was based on the data provided by the Nigeria Insurer Association (NIA) for the period 2013 to 2016. The model applied by the scholar was the Negative Binomial-Generalized Pareto Distribution (NB-GPD) and the Monte Carlo method for the modeling and simulation respectively. The expected total claims and the capital requirement were predicted. The essence of the study was to come up with a fair Reinsurance Price for large claims of Motor Insurance. The scholar concluded that the EVT is a powerful approach for the tail modeling of huge insurance losses. According to the scholar, the NB-GPD fit the tails of the motor claims well, as evidenced by the Cramer-von Mises and Anderson Darling tests. Based on a simulated yearly loss distribution, the predicted annual total claims for the motor insurance business were calculated.

The current research is primarily focused on modeling and predicting insurance claim amount based on the information available. A review of the literature revealed that most scholars analyzed only a single class of insurance businesses. This research attempts to bridge the gap by investigating all the non-life insurance classes of business in Nigeria. Furthermore, this is the first study to apply the Tree-Based models to analyze insurance data in the Nigerian insurance market.

2.3.2 Historical Development of Insurance in Nigeria

Before independence, the colonial era established the foundation for insurance activities in Nigeria. Through the inception of inter-regional trade, European enterprises provided services. These enterprises obtained insurance licenses from regulators abroad because they needed to limit risk on a grassroots. In the post-independence era, major European firms needed to maintain their dominance

University of Ghana http://ugspace.ug.edu.gh

in Nigeria. However, this resulted in the establishment of African branches for these businesses. Officially, the Royal Exchange Assurance Agency was the oldest insurance company in Nigerian history. The African Insurance Company Limited has the position for its indigenous variation. In 1958 it emerged and began the provision of insurance services. The following include other companies in operation at the time of the Royal Exchange Assurance Agency:

- Law Union and Rock
- BEWAC's Legal and General Assurance
- Patterson Zochonis [PZ] Liverpool

In an attempt by the Nigerian government to exercise its authority, the National Insurance Commission (NAICOM) was established. The Nigerian Insurance Market requires legislation that fits in with its immediate surroundings. In 1997 the NAICOM was created to monitor the country's insurance operations. The main insurance regulatory body in Nigeria is entrusted to NAICOM from then till now. The backdrop for expert and 'freshmen' policyholders is an overview of Nigeria's insurance history. Without our former colonial rulers, it is almost difficult to discuss the origins of insurance in Nigeria. Essentially, this offered the country's practice and operations platform. But we have seen constant improvement in the business in the last couple of years.

2.3.3 Brief Overview of the Insurance Industry

Africa is steadily moving toward a positive outlook, and a developing insurance industry contributes to the economy's development and prosperity. Savings and investment, employment creation, and expansion in capital markets and financial assets are all encouraged. In 2013, Africa saw a 12.8 percent increase in life premiums and a 2.1 percent increase in non-life premiums (Muguto, & Ubah,

University of Ghana http://ugspace.ug.edu.gh

2015). Through entry into the market, foreign investors have shown a strong desire to participate in the Nigerian insurance sector. Old Mutual and Sanlam from South Africa, as well as AXA from France, which bought a $246 million share in Mansard Insurance, are among them. The development of new insurance products in the booming mortgage and housing sector demonstrates progressivity. Despite the country's poor economic performance, the insurance business had a spectacular year in 2016. During the period, the industry gross premium, which is a crucial measure of industry growth, increased by almost 13 percent to over N326 billion, up from N289 billion in 2015 (NAICOM, 2016). The non-life sector maintained its market dominance, contributing to around 62 percent of the total gross premium, compared to 38 percent for the life business. This means that the Life business increased its percentage of overall industry gross premium by seven points, indicating a positive change from the previous year's contribution of 31 percent. During the same period, gross claims increased by 31 percent, more than doubling the prior period's 15 percent increase. During the year, non-life gross claims accounted for 54 percent of the total claims paid by the market, while the Life business side accounted for 46 percent. This occurrence signifies that insurance firms settled more claims over the period, and experience directly due to the Commission's continuous strengthened regulatory actions regarding genuine claims settlement, along with expanding consumer awareness. More performance analyses of the insurance industry would be provided in a detailed table in chapter four.

2.3.4 Classifications of Insurance Services

The two primary kinds of insurance service operations in Nigeria are life insurance and non-life insurance. The distinctions between the two are seen in the types of risks covered. Non-life policies cover against risks of loss resulting from the occurrence of specified events, whereas life insurance pays an amount if the insured

University of Ghana http://ugspace.ug.edu.gh

lives for a specified period or dies. Although both have shown stable growth in recent years, the latter has the biggest number of businesses. Non-life insurance companies are divided into six categories by the NAICOM: Motor, Fire, General Accident, Marine and Aviation, Oil and Gas, and Engineering. Motor insurance covers liabilities stemming from the operation of a vehicle. Fire insurance protects the insured organization by providing coverage i.e. in financial terms, in the event of property loss or destruction caused by fire, lightning, or explosion. In this case, fire refers to any type of blaze that can occur from any cause. General Accidents are insurance policies that typically cover unexpected, incidental, and abrupt loss or damage to property, as well as personal injury or death. General Accident Policies are divided into four categories: Property insurance, Insurance of the person, Pecuniary Insurance, and Liability Insurance. Marine insurance protects assets from hazards such as fire, abducting, and extreme weather conditions at the sea while airborne damages, as well as passenger and cargo liability, are all addressed in aviation insurance. Oil and gas insurance policies cover onshore risks such as refineries, depots, petrochemical plants, and tank farms, as well as offshore risks such as conglomerate and independent contract operators, as well as oil companies engaged in drilling rig operations, seismic operations, engineering, and construction, and a variety of other well delivery operations. Engineering Insurance is a type of insurance that protects against risks associated with engineering-related incidents. Engineering Insurance, to put it another way, covers the various types of risks associated with on-site building, such as the use of machinery and plants. As a result, non-life insurance offers a variety of coverage alternatives resulting in a diverse product line designed to meet the various risk needs of policyholders. This study would be based upon claims analysis for non-life coverage, as such, there are fewer comments on life insurance in this section.

University of Ghana http://ugspace.ug.edu.gh

Chapter 3

Methodology

The design of the research is given in this chapter. It covers an overview of the data, the empirical models, and the techniques of data analysis chosen for the research. As outlined in chapter one, the study attempts to find a suitable model that accurately predicts claims amount of non-life insurance claims of the Nigerian and Ghanaian Insurance market. The chapter goes on to the detailed statistical tools and procedures used for the investigation. The chapter highlights the essential theoretical notions that influence the author's implementation of selected methodologies in the broader academic research, as outlined by McCombes (2019).

3.1 Overview of the Data

The research makes use of secondary quantitative data on insurance claims amounts obtained from the Nigeria Insurer Association (NIA). The data available is from 2011 to 2018 with 10,017 observations for the Non-Life Insurance Claims. Some insurance businesses run composite lines, while most specialize in a single class, but specifically, the data collected were based on the following seven insurance categories: Motor Insurance claims, Fire and Special Perils Insurance Claims, General Accident Insurance Claims, Marine, and Aviation Insurance Claims, Oil and Gas Insurance Claims, Engineering Insurance Claims, and Miscellaneous Claims. On the other hand, the Ghanaian dataset comprises of motor vehicle insurance claims amount obtained from an insurance company in Ghana for the

University of Ghana http://ugspace.ug.edu.gh

period 2013 to 2016 with 5,495 observations.

3.1.1 Empirical Models

One of the most important variables to the Non-life insurance companies is the overall loss (L) for each component of risk exposure (e). The Overall loss (L) is the amount of loss for all insurance claims reported by the insured during the period of an insurance contract. To determine the price of insurance, some insurers choose the frequency-severity approach. Henkaerts et al. (2021) define the frequency of claims (F) as "the Number of claims (given as N) submitted for every unit of risk exposure (e), while the average amount per claim submitted, which is equal to the total loss amount (L) divided by the number of claims (N), determines the claim severity or amount of claims". Therefore, the price of an insurance contract (P) is given as:

$$P = E\Big(\frac{L}{e}\Big) = E\Big(\frac{N}{e}\Big) \times E\Big(\frac{L}{N} \mid N > 0\Big) = E(F) \times E(S) \tag{3.1}$$

The above price is determined where the premium components of the frequency and severity are assumed to be independent (Henkaerts et al., 2020). This implies that the actuary designs the models separately. But the severity model would only be developed with the dataset of policyholders that were paid insurance claims during the period. For the purpose of this study, the researcher adopts empirical models as in Quan (2019) to forecast claim amounts utilizing Tree-based models namely; CART, Random Forests, and Gradient Boosting under this investigation. Under the tree-based models, it is assumed that there exists a dataset containing the vector, k independent variables which are given by $X_i = (x_{i1}, x_{i2}, ..., x_{ik})$, sampled from space $X = X_1 \times .... \times X_k$ and a dependent variable, y_i, from sample space Y, for each of the N samples. The given sample dataset is denoted by (X_i, y_i) for $i = 1, ..., N$. The models adopted are discussed in detail in the

University of Ghana http://ugspace.ug.edu.gh

following subsection.

3.1.2 Classification and Decision Tree (CART)

The researcher would predict the amount or severity of the claim using the CART, following Quan, and Valdez (2018). The CART, developed by Breiman et al. (1984) requires recursive splitting to generate the regression trees that can forecast a continuous dependent variable for regression or a categorical dependent variable for classification. The study will employ the notation used by Hastie et al. (2009) and Quan and Valdez (2018). In this Machine Learning process, a regression (decision) tree, represented via $Z(x, \pi)$, is built by segregating a space that is linked to a particular response variable into Q disjoint regions $R_1, R_2, ..., R_q$ and then allocating a given constant c_q for every region R_q, for $q = 1, 2, ..., Q$. The regression tree can now be modeled as:

$$f(x_i \mid \pi_i) = \sum_{q=1}^{Q} c_q 1_{R_q}(x_i) \tag{3.2}$$

where $\pi = \{R_q, c_q\}_{q=1}^{Q}$ signifies the split with the allocated constants. Using the Classification and Regression Trees, constants are obtained through finding the minimum sum of squares errors cost function:

$$H(y, \hat{y}) = \sum_{i=1}^{N} (y - \hat{y})^2 \tag{3.3}$$

where $\hat{y} = \sum_{q=1}^{Q} c_q 1_{R_q}(x_i)$ is given as the dependent variable's estimated value. As clearly observed, the ideal value c_q is the mean of y_i of the domain R_q given as:

$$c_q = \text{average}(y_i \mid x_i \in R_q) = \frac{1}{N_q} \sum_{x_i \in R_q} \hat{y}_i \tag{3.4}$$

Where N_q stands for number of samples in the domain or portions of R_q. The portions or sections in the regression decision trees are created through the al-

University of Ghana http://ugspace.ug.edu.gh

gorithm, repetitive binary splitting. According to Quan and Valdez (2018), the initial stage in this technique is to choose one independent variable $X_{.m}$ that best splits the data samples into two subdomains or subsets. For instance $R_1(m,n) = \{x_i | X_{i.m} < n\}$ and $R_2(m,n) = \{x_i | X_{i.m} \geq n\}$ in this case the continuous independent variable. Hence the division is determined as shown below for any m and n. Thus:

$$\underset{m,n}{\operatorname{argmin}} \sum_{i:x_{i\in R_{1(m,n)}}} (y_i - \hat{c}_{R_{1(m,n)}})^2 + \sum_{i:x_{i\in R_{2(m,n)}}} (y_i - \hat{c}_{R_{2(m,n)}})^2 \tag{3.5}$$

The algorithm then searches for another independent variable with the optimum separation into two subgroups, repeating the procedure until the lowest amount of data in the section or other predetermined threshold is achieved. The outcome is T_0, a completely developed tree with several terminal sections, which could result in fitting and unwanted complexity of the model. This issue can be overcome by pruning the fully developed tree T_0 via cost-complexity pruning (more detail in subsection 3.4 of this chapter).

The loss in the region R_q can be calculated using equations (2) and (3):

$$L_q(T) = \frac{1}{N_q} \sum_{x_i \in R_q} (y_i - \hat{c}_q)^2 \tag{3.6}$$

For any root node, given as $T \subset T_0$represent the amount of terminal nodes in the subtree given as $\mid T \mid$. Thus, to enable control of the amount of terminal sections or portions, an alteration parameter is introduced given by $\alpha \geq 0$ so that the loss function can be redefined as:

$$C_\alpha(T) = \sum_{q=1}^{|T|} N_q L_q(T) + \alpha |T| \tag{3.7}$$

However, the loss function should always be chosen based on the distribution of

University of Ghana http://ugspace.ug.edu.gh

the data. If the data is normally distributed, the MSE is appropriate; however, if the data is distributed according to a Poisson distribution, the MSE may not be the right approach. But in this research, no distribution assumptions were made because the applied models does not require distribution assumptions.

The optimization parameter significantly penalizes a huge range of regions, going by the above cost function. The aim is to determine the $T_\alpha \subset T_0$ subtrees for each α and pick the one that gives the minimum $C_\alpha(T)$ Additionally, equivalent to the normalization parameter in a penalized model of regression, the tuning parameter α determines the trade-off between goodness-of-fit towards the dataset and the depth of the tree. The tree tends to be smaller if there are higher values of α. But when $\alpha = 0$, then results in a full-grown tree T_0. The tuning parameter α is estimated using K-fold cross-validation.

3.1.3 Random Forest

In this model there exist a collection of datasets given by $T(x, \Theta_b)$ which are ensembles of unpruned regression trees collectively referred to as Random forests. The b in the data is given as $b = 1, ..., B$. The B indicates the bootstrap samples created from the existing training dataset. This study will also follow the same assumption as Breiman (2001) which points out that $\{\Theta_b\}$ can be assumed to be as a set of randomly distributed random vectors that are independently and identically distributed. Therefore, the definition of the model aimed at the dependent variable is the average of all the regression decision trees in the Random Forest, where B is the whole number of the bootstrap observations. The random forests is given by:

$$f_B(X|\Theta) = \frac{1}{B}\sum_{b=1}^{B} f(X|\Theta_b) \tag{3.8}$$

University of Ghana http://ugspace.ug.edu.gh

Random forests lower the relationship between CART even more by choosing the optimum split out of a random group of independent variables at each tree node. Cross-validation can be used to determine the best size for the randomly selected subset of explanatory variables; some researchers use a general rule like the square root of the aggregate sum of regressors, $\sqrt{p}$ within the training data sample (Quan, 2019). The average predictions over several CART trees are projected to result in a reduced variance than one single CART tree. A significant randomized set of independent variables might enhance the predictive accuracy of any single tree, but at the same time, they may also strengthen the relationship between trees and then invalidate any benefit as a result of the average of numerous predictions. The bootstrap data resampling used to train every tree also adds to the variation across the trees. Random forest accuracy is determined by the strength of every single tree and the degree to which they are interdependent. Quan (2019) posits a principle of the Law of Large Numbers. The Law provides that when $B \to \infty$, we will obtain:

$$E_{x,y}(y - f_B(x|\Theta))^2 \to E_{x,y}(y - E_\Theta f(x|\Theta))^2 \tag{3.9}$$

In this case, Quan, and Valdez (2018) outline that the generalization error of random forests has a finite value, which explains the reason they rarely overfit when additional trees are introduced.

3.1.4 Gradient Boosting

Gradient-boosted decision trees are a machine-learning approach to optimize the explanatory power of a model via continuous learning stages. Each stage of the decision tree entails altering the coefficient values, weights, or bias values used to estimate the optimal model of each of the response variables to decrease the prediction error (deviation measurement between the expected values anticipated

University of Ghana http://ugspace.ug.edu.gh

and observed). In the Gradient Boosting, the regression decision tree is built or created by the previously developed tree's information so that every regression tree subsequently concentrates upon learning about the residues from prior trees (Quan & Valdez, 2018). The gradient is the stepwise adjustment performed at every stage of the process; boosting is a way to accelerate predictive accuracy improvements to a reasonably high level. According to Breiman (2001), the result of the tree would be a set of J regression trees $T(X;\Theta_J)$, for $j = 1, ..., J$. The GBRT approach would, therefore, be mathematically represented as a sum of the trees

$$F_J(X|\Theta) = \sum_{j=1}^{J} T(X;\Theta_j) \tag{3.10}$$

Where $\Theta_J = \{R_{iJ}, c_{iJ}\}_{i=1}^{I_j}$ and $T(X;\Theta_J)$ are corresponding models developed by the trees. The number of iterations in the procedure is called J. In the process, the ideal solution of Θ_J for each stage can be determined below:

$$\hat{\Theta}_J = \underset{\Theta_J}{\operatorname{argmin}} \sum_{i=j}^{N} L\Big(y, F_{j-1}(X_i|\Theta) + T(X_i|\Theta_s)\Big) \tag{3.11}$$

It, therefore, implies that:

$$F_j(X_i, \hat{\Theta}) = F_{j-1}(X_i|\Theta) \sum_{i=j}^{I_j} \hat{c}_{ij} 1_{R_{ij}}(X_i) \tag{3.12}$$

where $\hat{c}_{ij} = \underset{c}{\operatorname{argmin}} \sum_{i \in R_{ms}} L\Big(y, F_{j-1}(X_i|\Theta) + c\Big)$ It facilitates the regression tree, whereby the present residuals are optimally forecasted as $y_i - F_{j-1}(X_i|\Theta)$, under the SSL function, and $\hat{c}_{ms}$ is the average of the residuals.

3.2 Tuning Parameter Method

Many algorithms have parameters that, although crucial, cannot be predicted straightforwardly from a dataset. Hence, such tuning parameters are significant

University of Ghana http://ugspace.ug.edu.gh

because they frequently regulate the model's complexity and, as a result, any bias-variance trade-off that could be made is also affected (Kuhn, & Johnson 2019). In machine learning algorithms (such as CART), larger trees take higher variance with low bias, whilst smaller tree takes lower variance with high bias. To avoid overfitting, the number of regions represented by M penalizes the regression tree's performance as follows:

$$\sum_{i=m}^{M} \sum_{i:x_i \in R_m} L(y_i, \hat{y}_{i_{R_m}}) + M.cp. \sum_{i:x_i \in R_2(m,n)} L(y_i, \hat{y}_{i_{R_m}}) \tag{3.13}$$

In the above equation, the first component evaluates the goodness of fit, while the second part is a penalty that measures the complexity of the regression tree. To maintain the best trade-off and achieve a great model performance, a tuning parameter method needs to be applied. The most common methods for parameter tuning are the Grid Search, Random Search, and Automatic. For this study, the Grid search method of tuning (e.g. the Bayesian optimization) is selected since it is simple to apply but still being an effective method of tuning (Henkaerts et al., 2020).

3.2.1 Grid Search

One of the most common methods of hyperparameter optimization is the Grid search. In other words, David (2020) defines a Grid search as a process for determining the optimum value for a particular model by performing hyperparameter tuning. It's a time-consuming search on a manually defined subset of hyperparameter space. Cross-validation is commonly used to find the optimal combination of hyperparameters. Setting boundaries and discretization are required for obtaining the subset, which necessitates model competence. The R software would be used to perform the Grid search using the steps followed by Joseph (2018).

University of Ghana http://ugspace.ug.edu.gh

3.2.2 Cross-Validation

A cross-Validation is an efficient tool for monitoring the performance of a model, especially in a situation where the problem of overfitting needs to be addressed. It can also be used to determine the hyperparameters of a model. Cross-Validation is a common and simple tool used for quantifying prediction error (Arlot, & Celisse, 2010). The detailed procedure of Cross-validation can be seen in (Arlot, & Celisse, 2010). This study uses a k-fold method of cross-validation. The$k$-fold cross validation fits the model with a portion of the existing data and then tests it with the other half. In other words, the existing dataset is divided into k subsamples of equal size. In each iteration, one of the k subsamples is chosen as a validation dataset, while the remaining $k-1$ subsamples are utilized to construct and estimate the model's prediction error. Having performed the procedure for $k = 1, 2, ..., K$, this technique now calculates the mean of the estimation error from k validation datasets. Assuming $k : \{1, ..., N\} \rightarrow \{1, ..., K\}$ is an indexing factor which specifies the segment to what observation i is randomly assigned. We denote $f^{-k}(x)$ to be the fitted function, calculated with the kth part of the data censored. In this case, the prediction error estimates from cross-validation are used (Hastie et al., 2009). It is given as:

$$\text{Cross Validation } (\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i)) \tag{3.14}$$

As stated in previous paragraph that if $k = \{1, ..., N\} \rightarrow \{1, ..., K\}$, then it implies that the k-fold cross-validation yields the leave-one-out cross-validation (). The size of the training set and the number of splits are both fixed if the number k is chosen. Since a bigger k leads to a larger training set, hence the k-fold cross-validation bias reduces with k. While the variance of k-fold cross-validation grows with k because a bigger k leads to a higher number of validation procedures. The value k is usually within 5 and 10 in practice. According to Quan

University of Ghana http://ugspace.ug.edu.gh

(2019), 10-fold cross-validation has been proved to produce good point estimation and model selection. In summary, the steps involved to carry out cross-validation procedures are as follows:

- Divide the datasets F into 6 disjoint separated subsets $F_1, ..., F_6$
- For $k = 1, ..., 6$

  Leave out F_1 as a test dataset

  Perform 5-fold cross-validation on $F \backslash F_k$

  Train the model on $F \backslash F_k$ using the optimal parameters

  Evaluate the performance of the model on F_k The result would be Six (6), optimal models, for each procedure

3.3 Performance Measures of the Tree-Based Models

Vrieze (2013) defined a model as a mathematical representation of the correlation between measurement properties. A model is a framework that depicts how a variable behaves given a specified phenomenon. In machine learning and other standard statistical methods, performance measures are widely employed for model evaluation. The purpose of this evaluation is to determine which of the model best fits the data under investigation. There is no single great predictor which can under any circumstances be used to evaluate predictive performance and accuracy. However, the one that best performs in terms of specified performance indicators is finally selected from several trained models with various sets of parameters and hyper-parameters (Vaid, 2019). For this study, we use a few standard measures to make a valid assessment of models. They are; Coefficient of determination (R^2), Mean Absolute Error (MAE), and Mean Squared Error

University of Ghana http://ugspace.ug.edu.gh

(MSE).

3.3.1 Coefficient of determination (R^2)

The coefficient of determination was introduced by Sewall Wright and was developed in 1921. In a correlation analysis framework, Enders (2020) describes the coefficient of determination (referred to as R^2) as a metric that evaluates a model's ability to interpret or predict a result. R^2 denotes the percentage of dispersion of the dependent or response variable (Y) explained or predicted by linear regression and the independent or explanatory variable (X). Suppose there is a dataset that has n number of values given by $y_1, ..., y_n$ collectively known as y_i, each related with a fitted value $\hat{y}_1, ..., \hat{y}_n$. The residuals would be given as $e_i = \hat{y}_1 - \hat{y}_n$. If $\bar{y}$ is the mean of the observed dataset given by $\frac{1}{n}\sum_{i=1}^{N} y_i$. The variation of the data can then be measured with two sums of squares given by $\sum_{i=i}^{N}(y_i - \bar{y})^2$ and residual sum of squares given by $\sum_{i=i}^{N}(y_i - \hat{y})^2$ Therefore the R^2 would then be mathematically represented as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \frac{1}{N}\sum_{i=1}^{N} y_i)^2} \tag{3.15}$$

R^2 provides information regarding a model's goodness of fit. Its values run from 0 to 1 and represent how strong the correlation predictions estimate the actual data points. As a result, an R^2 of 1 shows that the correlation predictions accurately fit the actual data. In other words, a greater R^2 is better.

3.3.2 Mean Absolute Error (MAE)

The mean absolute error (MAE) is a statistical approach that measures the difference in errors amongst paired observations describing an identical phenomenon. It involves evaluations of predicted and observed results. It can be mathemati-

University of Ghana http://ugspace.ug.edu.gh

cally represented as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}_i| = \frac{1}{N}\sum_{i=1}^{N}|e_i| \tag{3.16}$$

As a result, it's an arithmetic mean of absolute errors. y_i is the predicted value and $\bar{y}$ is the actual value. The interpretation of MAE implies that a lower MAE is better.

3.3.3 Mean Squared Error (MSE)

MSE is a statistical method for calculating the mean of the squared errors, or the mean squared difference between the actual and estimated values. The MSE compares the difference between the model's predicted values and the actual values, all squared, and then averaging it across the entire dataset. It is mathematically given as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y - \hat{y})^2 \tag{3.17}$$

Its interpretation is that a lower MSE is better.

3.4 Statistical Computing Environment

The researcher employs R software in conducting the tree-based analysis of the dataset under investigation. Specifically, libraries such as rpart, randomForest, gbm, caret, and ranger in the R environment, offer a complex set of actuarial and statistical tools which aid in fitting the models to data of the research. The tools would allow the researcher to analyze the predictive performance of the models to enable comparison and select the best.

University of Ghana http://ugspace.ug.edu.gh

Chapter 4

Data Analysis and Discussion

This chapter entails the detailed analysis of data, the results, and a discussion of the study. Regarding the methods discussed in the previous chapter, this chapter describes the model fitting of the Tree-Based Machine Learning (ML) Algorithms considered for the prediction of insurance claims amount based on the Nigerian and Ghanaian Insurance data.

4.1 Data Description

As briefly stated in the previous chapter that this research makes use of secondary qualitative data on insurance claims amount obtained from the Nigeria Insurer Association (NIA). The data entails large claims amounts that were settled and paid by the non-life Insurance class of business in Nigeria. The available data covers the years 2011 to 2018. The major categories of Insurance covers in Nigeria are Motor, Fire and Special Perils, Marine and Aviation, Engineering, Oil and Gas, and Miscellaneous Insurance claims. The data is further described in detail in terms of its basic statistics and limitations.

4.2 Descriptive Statistics

A detailed statistical summary of the dataset would provide useful information to facilitate the analysis of the data. The minimum, maximum, mean, standard deviation and skewness of the datasets are provided in table 4.1.

University of Ghana http://ugspace.ug.edu.gh

Sample Size	Min	1stQ	Mean	Median	3rdQ	Max	Std Dev	Skew
10017	1300000	4345000	16976315	5985000	10810000	7876000000	144235790	54.15

Table 4.1: Summary Statistics of Insurance Claims Amount

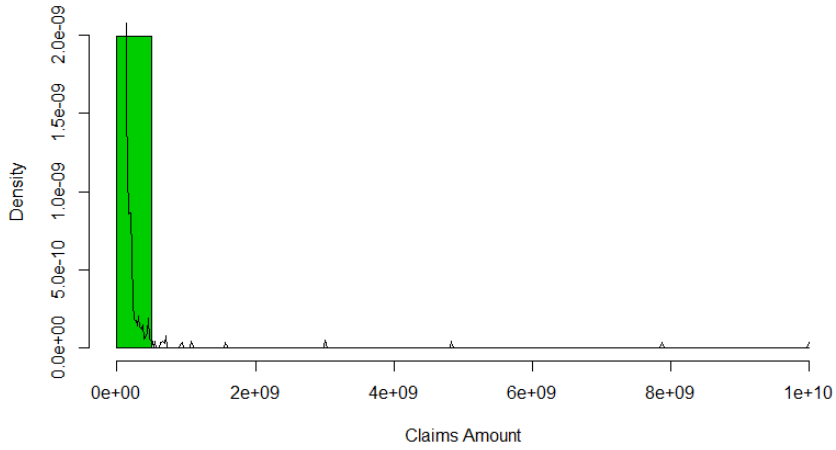


Figure 4.1: Histogram of the Claims amount

Table 4.1 show that there is a total of 10,017 observations in the data. The statistics show the data is highly skewed to the right throughout all the years. This is attributed to the fact that the mean is greater than the median. This can be graphically represented by figure 4.1.

Figure 4.1 is a histogram plot of the whole dataset. The plot shows that there are very high insurance claims amount that does not occur frequently. Table 4.1 show that the standard deviation of the data is very high and above the mean. This implies that the data is spread out. The dataset is further broken down to provide the summary statistics of the various insurance categories given in table 4.2. From table 4.2 we can see that type of insurance cover with the maximum claims amount is the Oil and Gas Insurance Cover which occurred in 2018. The second type of insurance cover having the highest insurance claim amount is the General Accident.

The result in table 4.2 also shows that the claims amount for all insurance categories is right-skewed. Because the mean claims amount is above the median

University of Ghana http://ugspace.ug.edu.gh

Insurance Cover	Sample Size	Min	1stQ	Mean	Median	3rdQ	Max
Engineering	361	3530235	4946490	13041178	7105365	11344926	274455149
Fire and Special Perils	2476	1299747	4890000	19573028	7638840	15710000	1571330767
General Accident	1804	3000000	4267000	13892609	5507888	9000000	7875593935
Marine and Aviation	956	3081250	4790402	18313004	7625298	15576702	631343671
Miscellaneous	730	2062455	4500000	14986862	6650564	12383773	491414142
Motor	3017	3035975	4030000	5841740	4738500	6258625	51810400
Oil and Gas	673	1849000	7419000	67970000	16250000	43200000	9999000000

Table 4.2: Summary Statistics of Claims Amount: Insurance Cover

in all the categories of insurance. The summary statistics are also provided for the types of entities that purchase insurance policies in the Nigerian Insurance market.

Type of Entity	Sample Size	Min	1stQ	Mean	Median	3rdQ	Max
Corporate	8915	1299747	4386000	18030000	6132000	11430000	9999360000
Public	345	3249000	4463600	12441302	5717600	9656410	227886698
Individual	726	1849182	4031736	6450139	4805735	6514786	107301681
Others	31	3500000	4707500	12261327	7020000	18320358	57406267

Table 4.3: Summary Statistics of Claims Amount: Type of Entity

From table 4.3, we can see that the mean is above the median meaning that the data is rightly skewed for each type of entity. Table 4.3 further shows that the corporate entities have the highest-paid insurance claims amount. This can give us an insight into the penetration of the insurance business into the Nigerian insurance market. Additionally, the dynamics and types of risks faced by the corporate entity differ from the types of risks faced by ordinary individuals in terms of the insurance.

Variable	Type	Description
Year	Integer	min: 2011; max: 2018
Type of Loss	Categorical	16 levels; accident, damage to property, illness, death, etc.
Type of Entity	Categorical	4 levels; corporate, individual, public, etc.
Type of Insurance cover	Categorical	7 levels; fire, motor, oil and gas, etc.

Table 4.4: Description of predictive (explanatory) variables of Nigerian Dataset

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

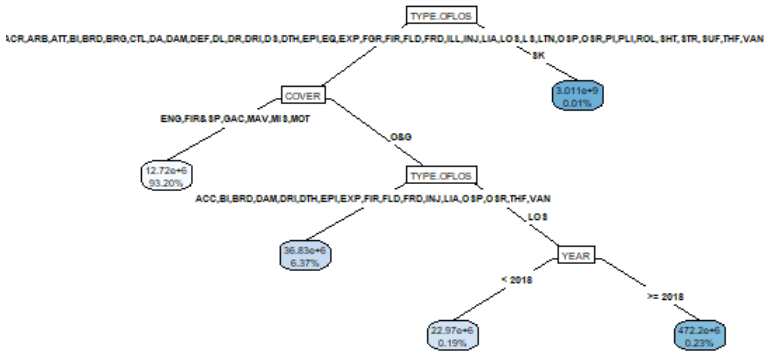


Figure 4.2: Regression tree plot

4.3 Fitting the Machine Learning Models: Nigerian Data

To fit the Tree-Based ML models, the dataset was split into training and testing. The proportion of the split in this research was selected at random. 70 per cent of the data for training and fitting, and then the remaining 30 per cent. The training data had a total of 7,012 observations and the testing data had a total of 3,005 observations. In this section, the CART, RF, and GB models are fitted to the training data by tuning the parameters of the selected ML Algorithms. And then finally, we compare the performance of the models used in the analysis to test the predictive accuracy on both the training and testing datasets.

4.3.1 The Classification and Regression Tree (CART)

The first step to building the CART model is to fully grow the tree on the predictors X_j and then cut the point called g in a form that splits the dataset into two distinct regions $m_R = \{X|X_j > s\}$ and $m_L = \{X|X_j \leq s\}$. The results are given figure 4.2.

Figure 4.2 is the plot of the regression tree results after tuning. The first split

University of Ghana http://ugspace.ug.edu.gh

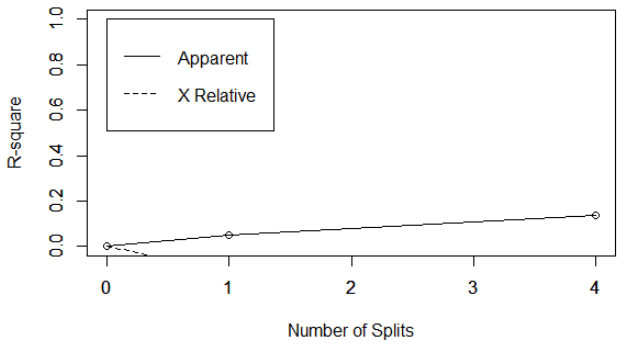


Figure 4.3: The Optimal number of splits

uses the type of loss and the remaining split uses the year and cover. As seen from the diagram, the variable, type of loss is the most important variable for the regression tree. If observations meet the splitting criteria, then the model assigns them to the left node, if not, the model assigns them to the right. In each node, there is information about the number of predictions and observations. The regression tree plot, further shows that the darker blue color implies a larger value of prediction. For this dataset, we set the pruned tree to 4 as the minimum number of observations for a split to be carried out. But the default minimum number of this tuning parameter in the rpart package is 20. This is known as the "minsplit". This has been adjusted to 8 based on the R-Squared of the fitted model. Figure 4.3 shows the ideal number of splits to be used for the final tree. From the plot, we can see that the optimum number of the split is eight. This is selected based on the number of the split with a higher value of the R-Squared.

Another important tuning parameter to prune the CART is to control the complexity of the model using the complexity parameter (CP). The CP method is used to perform the tree pruning so that the best tree can be determined. The manually specified subset of parameter space is the combination of the complexity parameter ranging from 0.001 up to 0.01 based on the increase of 0.01 of 10. The optimal parameter was selected using the RMSE. This procedure uses a grid search using 10-fold repeated cross-validation.

University of Ghana http://ugspace.ug.edu.gh

CP	RMSE
0.001	60096761
0.002	60088349
0.003	60056193
0.004	60013969
0.005	59823627
0.006	59589023
0.007	59589502
0.008	59562914
0.009	59624916
0.010	59647192

Table 4.5: Grid Search Result

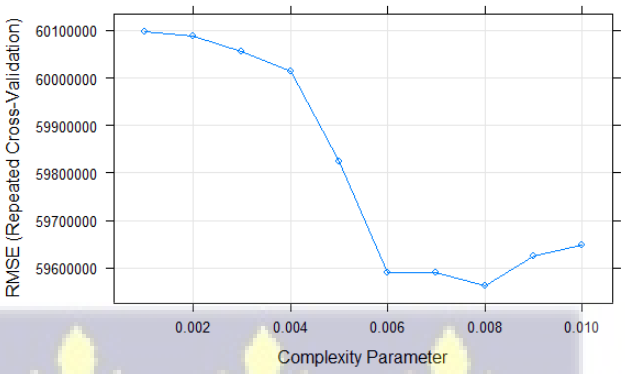


Figure 4.4: Plot of the Complexity Parameter against the RMSE

The CP value used to choose the final model is 0.008 on 7,012 samples with 4 predictors. The optimum CP was selected based on the lowest Root Mean Squared Error (RMSE) of 59562914. Figure 4.4 is a plot of the CP against RMSE.

The plot that shows the various points at which the RMSE increases or decreases as the CP parameter is tuned and validated on 10-fold cross-validation. From the plot, the RMSE moved gradually from 0.001 to 0.004 and then had a sharp decrease to 0.006. From the plot, we can see that the optimum CP is 0.008. Finally, Table 4.6 shows the important variables that were of significance in building the CART.

Variable	Overall
Cover	0.041
Type Of Entity	0.003
Type of Loss	0.282
Year	0.103

Table 4.6: Variable Importance

University of Ghana http://ugspace.ug.edu.gh

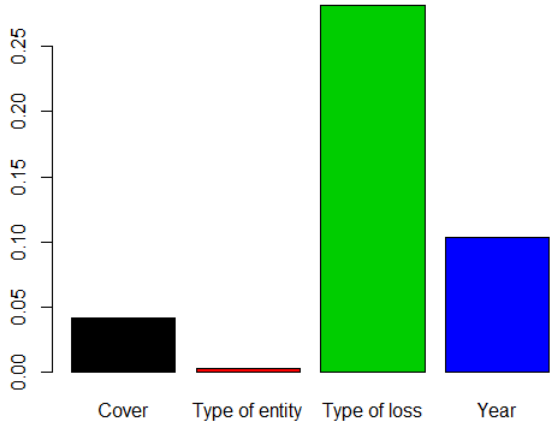


Figure 4.5: Plot of Variable Importance (CART Model)

The variable of importance explains the contributions of the variables by ranking them based on the highest decrease in error. Therefore, based on the criteria, table 4.6 shows that the most important variable is the type of loss. This implies that the variable contributes most to the model. Hence more important in determining the claims amount. The second most important variable in building the CART model is the year followed by the cover. Lastly, the variable with the least importance is the type of entity.

Figure 4.5 is a plot of the variable importance for the CART model.

Node	Obs	Wt	Dev	Yval	Complexity
4	6535	6535	6.535382e+18	12716412	0.004836194
10	447	447	1.510085e+18	36832773	0.000296937
22	13	13	9.873040e+15	22966147	0.008000000
23	16	16	2.142070e+19	472186450	0.008000000
3	1	1	0.000000e+00	3010669369	0.008000000

Table 4.7: Final Results of the fitted CART tree (NG Data)

Similarly, table 4.7 above is a numerical representation of Figure 4.2 for the final fitted CART. The Yvals are the leaves of the tree or the predicted amounts. The claims amount is then predicted based on the final fitted CART Model using both the training and testing datasets.

University of Ghana http://ugspace.ug.edu.gh

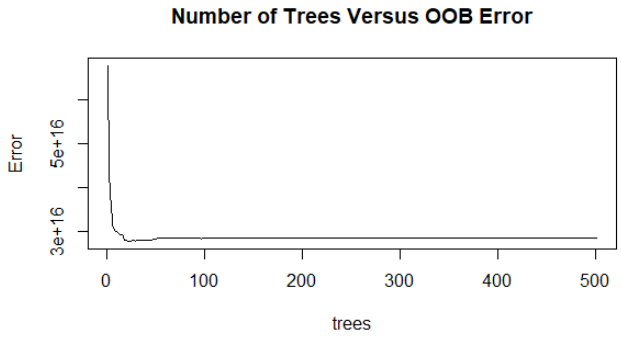


Figure 4.6: Optimum no. of Trees vs OOB Error

4.3.2 Random Forest

Due to the large number of regression trees required in a random forest, there is a need to focus on some essential tuning parameters (Quan, 2019). The tuning parameters of the random forest are explained by the out-of-bag (OOB) error rate and the number of trees. The default number of trees to build the random forest model in R statistical software is 500 Trees. In this thesis, we maintained the default so that we can later decide to increase or decrease this number to get the optimum for the final model. Figure 4.6 is a plot of the out-of-bag (OOB) error based on the number of trees in bagged trees. However, the prediction error computed from this estimation is termed the OOB error. From Figure 4.6, it can be seen that the OOB error subsequently reduces and normalizes after about 200 trees. We, therefore, tune the tree to fit our dataset as the OOB error of the model tends to reduce at this point.

The plot further displays the correlation between the OOB error and the number of trees. This implies that when there are few trees, there would be higher OOB error and vice versa. Quan (2018) outlined that this is a normal procedure because the OOB error rate tends to reduce as the number of trees increase. Therefore, the decision of the random forest on the optimum number of trees is dependent on this reduction. Our results are in line with Quan (2018), from the plot we can see that a few numbers of tree results in higher OOB error. The next step is to

University of Ghana http://ugspace.ug.edu.gh

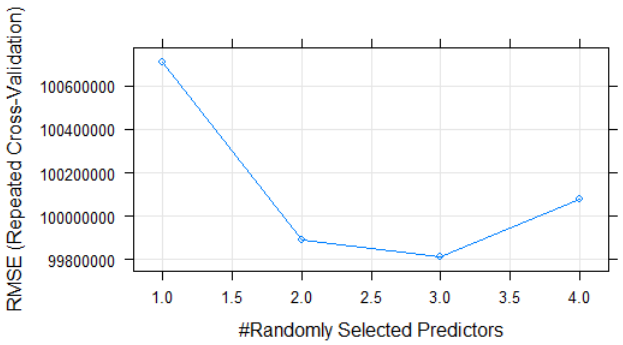


Figure 4.7: Randomly Selected Predictors

tune the Random forest model to determine the optimum mtry parameter to use based on 10-fold cross-validation repeated one time. The result of this validation would be used to select the optimum number of variables tried at each split for the final random forest model. Table 4.8 is the results of the cross-validation of 7,012 samples with 4 predictors.

mtry	RMSE
1	100711914
2	99891698
3	99810019
4	100080028

Table 4.8: Optimum no. of mtry

We can see that based on table 4.8, the optimum mtry to select in tuning the final Random forest model is 3.

Figure 4.7 shows that the RMSE increases or decreases depending on the complexity of the mtry parameter. It is worthy to note that, the final aspect of the RF model is the relevance of the variables in determining the appropriate number of trees for the model. The two standard measures of variable importance selected by the RF model are the "%IncNodePurity" and the "%IncMSE". These measures are useful in determining the optimum number of trees for the random forest model.

The variable with higher values of %IncNodePurity (or lowest node impurity) implies that the variable has a higher level of importance. Therefore, from table

University of Ghana http://ugspace.ug.edu.gh

Variable	IncMSE	IncNodePurity
Year	3.050761e+14	1.896689e+19
Type Entity	3.053951e+15	5.216019e+18
Type of Loss	1.316469e+13	1.902051e+19
Cover	1.482856e+14	7.227650e+18

Table 4.9: Variable Importance

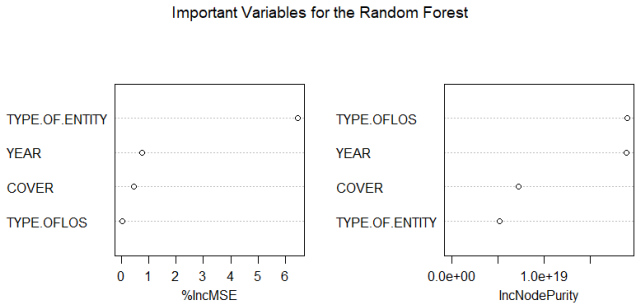


Figure 4.8: Variable Importance Plot

4.9, the variable type of loss has the highest value of 1.902051e+19, followed by the variable year with a value of 1.896689e+19. The type of entity and cover has a lower significance with values 5.216019e+18 and 7.227650e+18 respectively. On the other hand, a higher %IncMSE implies higher variable importance. Hence, from table 4.7 we can see that the variable Type of entity has the highest IncMSE with a value of 3.053951e+15 followed by the variable year with a value of 3.050761e+14. For measure, the Cover and type of loss were of less significance with a value of 1.316469e+13 and 1.482856e+14 respectively.

Figure 4.8 is a plot of the important variables for the RF model based on the %IncNodePurity and %IncMSE. Having observed the two popular measures of variable importance for RF, some scholars are of the view that %IncMSE gives a better and precise measure of variable selection. This is attributed to the fact that the %IncNodePurity may result in a biased variable selection process Inherited by the regression tree (Quan, 2018).

University of Ghana http://ugspace.ug.edu.gh

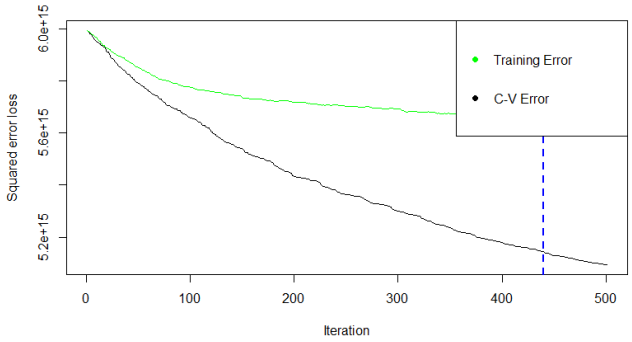


Figure 4.9: Selecting the best number of Trees

4.3.3 Gradient Boosting

The gradient boosted regression trees in this research are modeled based on a Gaussian loss function with 500 iterations. The result of this can be seen in figure 4.9. There were 4 predictors used by the model out of which all 4 had non-zero influence.

The plot show that the squared error loss on the number of iterations based on the training dataset on 10-fold cross-validation. As a result of this, we can say that the squared error loss tends to decrease as the number of trees increases but the optimum number of iteration is 440. The black lines represent the cross-validation error and the green line represents the Training error. It should be noted that the optimum number of iterations in the models was based on the value of shrinkage 0.01, an interaction depth of 4, and a minimum number of observations in the tress' terminal node of 10. The small value of 0.01 for the shrinkage was used as recommended by most scholars because it improves the predictive accuracy of the model by taking smaller steps than larger steps. This implies that when there is an error, its consequences can be corrected easily. The final results of the gradient boosted model are best presented in terms of the relative influence of the explanatory variables.

From table 4.10 we can see that the variable, Type of loss is the most important

University of Ghana http://ugspace.ug.edu.gh

Variable	Relative Inf.
Type of Loss	57.47
Cover	26.77
Year	15.74
Type of Entity	0.02

Table 4.10: Relative Influence

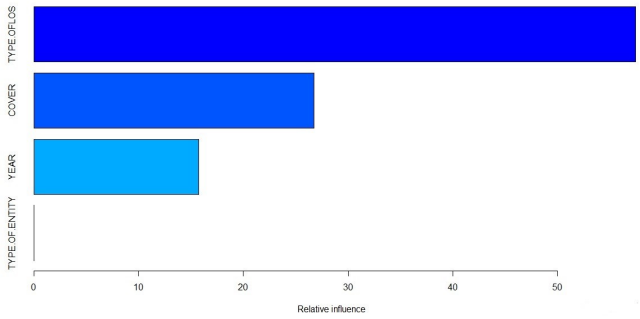


Figure 4.10: Relative Influence

variable used by the GB model with a relative influence of 57.47%, followed by Cover and Year with relative influence of 26.77% and 15.74% respectively. The variable with the least relative influence is the type of entity with a relative influence of 1.98%. This is graphically represented in figure 4.10.

As Quan (2018) rightly mentioned that the relative influence of each variable is further standardized so that the sum adds up to 100 percent, this implies that the larger values mean that it has a stronger influence on the response variable.

4.3.3.1 Partial Dependence Plots

The partial dependence plots show us the relative influence of each of the attributes under a particular categorical variable when other attributes are held constant.

From figure 4.11 it can be seen that under the variable cover, the attribute Oil and Gas (OG) contributes more to the claims amount when the others are held constant. This implies that the contribution of the attribute concerned, is far higher than the rest of the attributes.

University of Ghana http://ugspace.ug.edu.gh

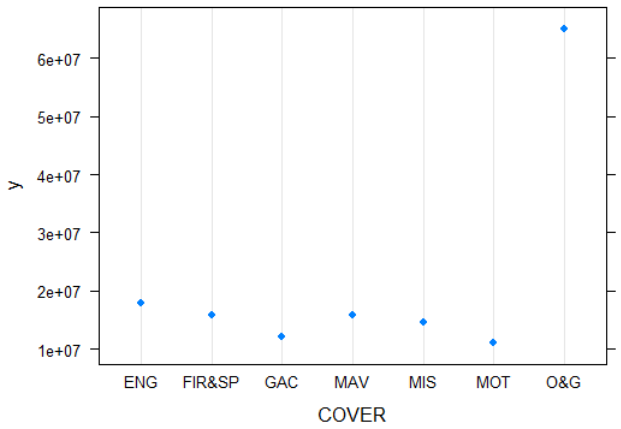


Figure 4.11: Partial Dependence plot Variable Insurance Cover

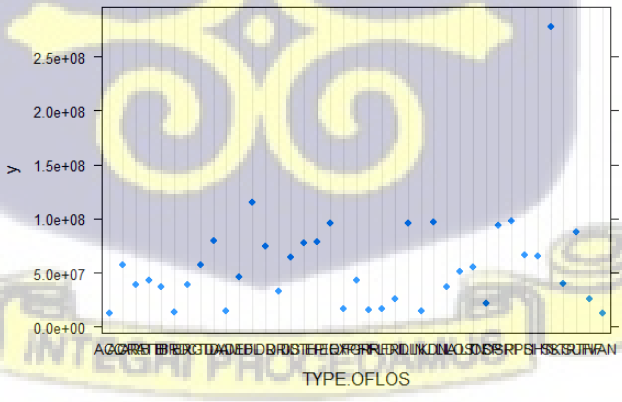


Figure 4.12: Partial Dependence plot Insurance Cover

University of Ghana http://ugspace.ug.edu.gh

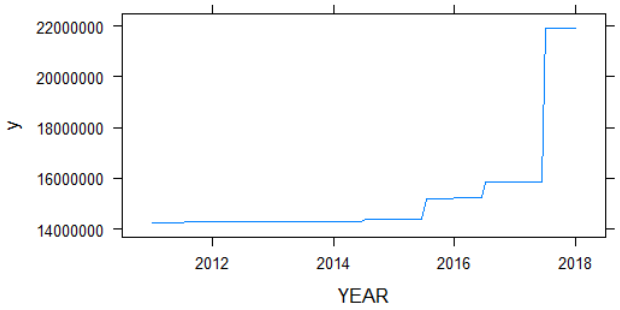


Figure 4.13: Partial Dependence plots Variable Year

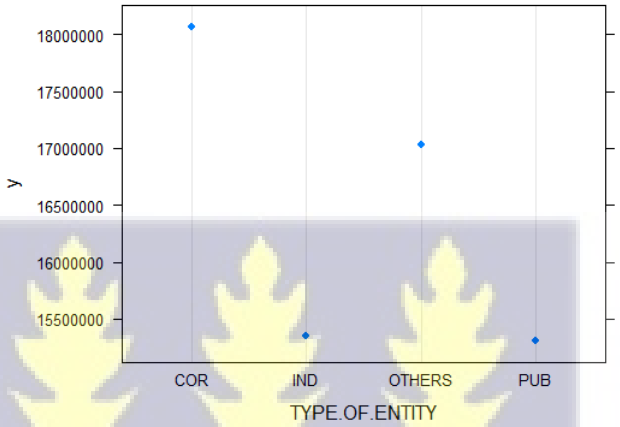


Figure 4.14: Partial Dependence plot Insurance Type of Entity

Figure 4.12 shows the partial dependence plots of the variable type of loss. Under this variable, we can see that the attribute sinking has the highest contribution in determining the claims amount when all other variables are held constant.

Figure 4.13 shows the relative importance of the year 2018. Meaning that it contributed more in determining the claims amount in the GB model. We can see that 2012 contributed the least to the model.

Even though we have seen that the type of entity was the variable with the least relative influence in building the GB model, figure 4.14 shows that under this variable the corporate category contributed the most in determining the claims amount while holding other variables constant in this category.

University of Ghana http://ugspace.ug.edu.gh

4.4 The Ghanaian Dataset

In this section, the results of the Ghanaian dataset are discussed. A description of the data and statistical analysis were carried out using the same Machine Learning Algorithms that were applied to the Nigerian Dataset.

4.4.1 Description of Data

The Ghanaian Dataset is a data of motor vehicle insurance claims collected from a major insurance company in Ghana which covers 4 years of data for the period 2013 to 2016. The data consist of seven (7) variables having six (6) categorical and one (1) continuous variable. The next section shows the description of the variables of the Ghanaian dataset.

4.4.2 Descriptive Statistics

Table 4.11 provides detailed statistics of motor vehicle insurance claims amount. The dataset has a total number of 5,495 observations. We can also see that the data is skewed to the right which explains why the mean is greater than the median.

Sample Size	Min	1stQ	Mean	Median	3rdQ	Max	Std Dev	Skew
5495	70	1189	7973.54	2975	7340	483602.60	17951.53	8.81

Table 4.11: Summary Statistics of Insurance Claims Amount (GH Data)

The dataset has a high standard deviation. This implies that the data is spread out. The dataset is graphically represented in Figure 4.15. The figure is a histogram plot of the dataset. The plot shows that there are relatively high insurance claims amount that does not occur frequently.

University of Ghana http://ugspace.ug.edu.gh

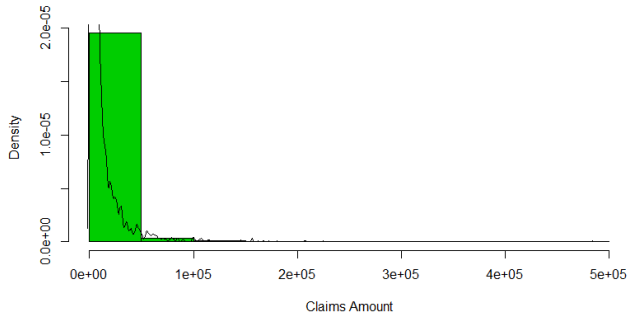


Figure 4.15: Histogram of the Claims amount(GH Data)

Variable	Type	Description
Insurance Type	Categorical	3 levels; Comprehensive, Third Party, Fire and Theft
Usage Type	Categorical	16 levels; Ambulance/Hearse, Corporate Buses, Hire Cars, Taxis, etc.
Purpose	Categorical	4 levels; Commercial Vehicles, Private Individual, Own Goods, etc.
Vehicle Age	Integer	min: 1; max: 49
Vehicle Make	Categorical	15 levels; BMW, Daewoo, DAF, Ford, Honda, Hyundai, KIA, etc
Category	Categorical	10 levels; Articulator, Bus, Lift Back, Motor Cycle-light, Saloon, etc.
Location	Categorical	10 levels; AS, BA, CR, ER, GT, NR, UE, UW, etc.

Table 4.12: Description of predictive (explanatory) variables of Ghanaian Dataset

4.5 Fitting the Machine Learning Models: The Ghanaian Dataset

In fitting the Ghanaian dataset, we follow the same procedures applied to the Nigerian Dataset. The Ghanaian Dataset has a total of 5,495 observations. The same ratio is also used to split the Ghanaian Dataset. In this section, the training dataset has a total observation of 3,846 representing 70% and the testing dataset has a total observation of 1,649 representing 30%. The same selected Tree-Based ML models were used to perform our analysis. We also compare the predictive accuracy of the models on the training and testing data.

4.5.1 The Classification and Regression Tree (CART)

The same procedure was used to grow the tree. For this dataset, the tree was grown based on 7 explanatory. The results of the final regression tree plot for the

University of Ghana http://ugspace.ug.edu.gh

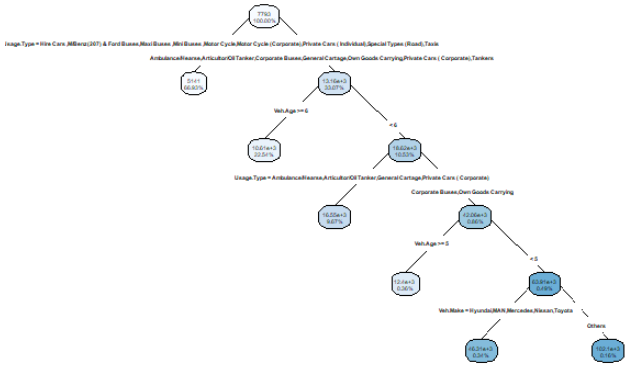


Figure 4.16: Plot of the Final CART Model GH Data

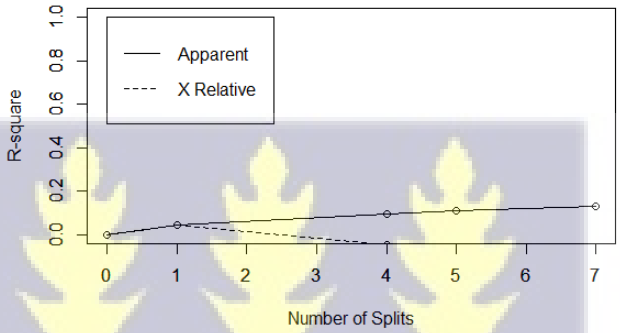


Figure 4.17: The Optimal number of splits

Ghanaian Dataset are given in figure 4.16.

From the figure we can say that the first split uses the vehicle make followed by the vehicle age. The next split uses Purpose and Category. From the figure, the darker blue color implies a larger prediction value on the dataset. Figure 4.17 is used to determine the optimum minimum number of splits measured by the highest R-Squared for the Ghanaian dataset. The result shows that the optimum number of the split is seven.

The next step is to determine the complexity parameter (CP) for the CART model fitted to the Ghanaian dataset. Table 4.13 shows the results of CP tuning based on the grid search.

From the table it can be seen that the optimum CP used for the CART model is

University of Ghana http://ugspace.ug.edu.gh

CP	RMSE
0.001	17255.47
0.002	17249.78
0.003	17210.01
0.004	17202.35
0.005	17246.33
0.006	17260.42
0.007	17245.12
0.008	17181.39
0.009	17141.25
0.010	17064.72

Table 4.13: Grid Search Result

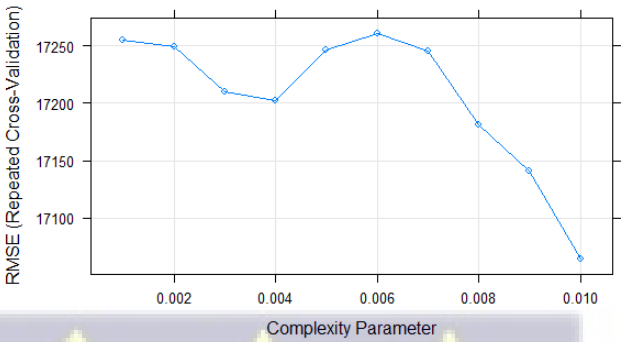


Figure 4.18: Plot of the Complexity Parameter against the RMSE

0.01 based on the lowest RMSE. The 10-fold repeated cross-validation was used based on the grid search.

Figure 4.18 shows the plot of the CP versus the RMSE. The plot moved gradually from 0.001 to 0.004 at a slow rate. The RMSE increased and then subsequently decreased to 0.01. After the Final CART model was fitted, the variable of importance was extracted. This can be seen in table 4.14.

Variable	Overall
Category	0.093
Location	0.016
Purpose	0.089
Usage Type	0.084
Vehicle Age	0.139
Vehicle Make	0.144
Insurance Type	0.000

Table 4.14: Variable Importance

The table show that the vehicle make contributed the most in determining the claim amount when compared to the other variables. The second and third most important variables that contributed to determining the claim amount are the

University of Ghana http://ugspace.ug.edu.gh

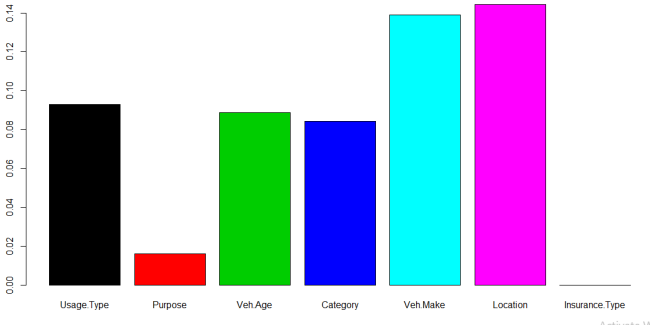


Figure 4.19: Plot of Variable importance for the CART Model

vehicle age and category. From the table, we can see that Insurance type has no influence in determining the claims amount under the CART model.

Figure 4.19 show the attributes and their influence on the claim amount.

Table 4.15 is a numerical representation of the final fitted CART model for the Ghanaian dataset.

Node	Obs	Wt	Dev	Yval	Complexity
2	2574	2574	236197485797	5140.994	0.004971524
6	867	867	260595007070	10608.509	0.007086679
14	372	372	382876024942	16545.715	0.007373834
30	14	14	4663967980	12398.816	0.010000000
62	13	13	36055699276	46308.424	0.010000000
63	6	6	181724234046	102056.288	0.010000000

Table 4.15: Final Results of the fitted CART tree GH Data

4.5.2 Random Forest

In fitting the Ghanaian dataset to the random forest model, we tune the parameter using the default value for the number of trees in R where (ntree = 500). Figure 20 is the plot of the OOB error. From the figure based on our estimation, it can be seen that the optimum number of trees required to build the random forest is 360 trees. At that point, the OOB error now stabilizes. This is used to fit the datasets to the final model of the random forest.

Additionally, figure 4.20 displays the correlation between the OOB error and the

University of Ghana http://ugspace.ug.edu.gh

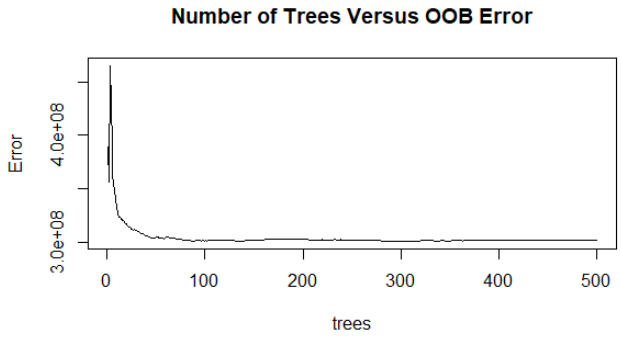


Figure 4.20: Optimum no. of Trees vs OOB Error GH Data

number of trees. The next step is to tune the RF model to determine the optimum mtry parameter to use based on 10-fold cross-validation repeated one time. The result of this validation would be used to select the optimum number of variables tried at each split for the final RF model on the Ghanaian Dataset.

mtry	RMSE
1	17585.16
2	17308.65
3	17143.63
4	17022.28
5	16934.83
6	16869.23
7	16814.95

Table 4.16: Optimum no. of mtry

From table 4.16 we can see that the optimum mtry to use in fitting the final RF model to select is 7 based on the lowest RMSE. This means that 7 is the optimum number of variables tried at each split for the final random forest model. Figure 4.21 is the plot of the cross-validation of 4,392 samples with 7 predictors. The plot shows the randomly selected predictors with the corresponding RMSE at each point.

Finally, after fitting the RF model, the variable of importance explains the predictive power of each attribute. Table 4.17 displays the results of the "%IncNodePurity" and the "%IncMSE". From the table we can see that the two most significant variables in determining the claims amount are the vehicle make and vehicle age having %IncMSE of 65565656 and 64935317 respectively. The next

University of Ghana http://ugspace.ug.edu.gh

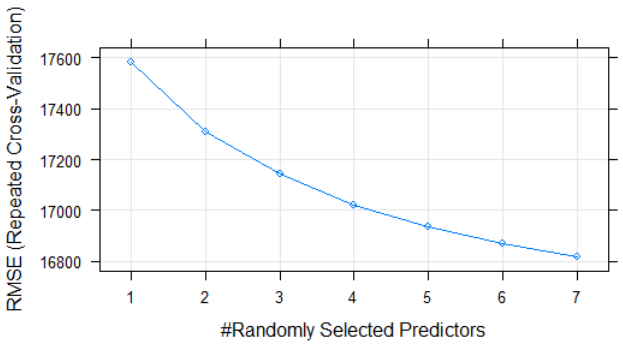


Figure 4.21: Randomly Selected Predictors

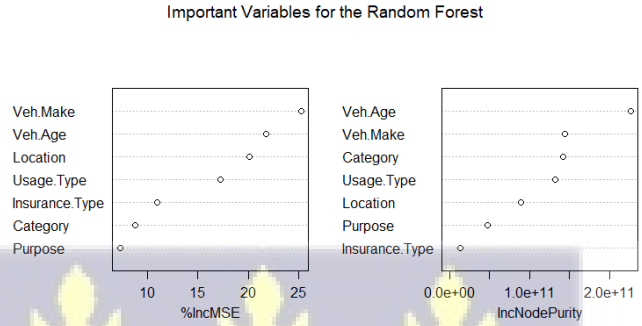


Figure 4.22: Variable Importance Plot

significant variables are the Usage Type, Location, and Category with corresponding %IncMSE of 50701576, 26730096, and 20720941 respectively. On the other hand, we can see that based on the %IncNodePurity the most significant variable is the vehicle age.

Variable	IncMSE	IncNodePurity
Insurance Type	4876002	13959155154
Usage Type	50701576	132836187330
Purpose	9890837	47326172496
Vehicle Age	64935317	226862234369
Vehicle Make	65565656	144494440894
Category	20720941	141869306583

Table 4.17: Variable Importance

Lastly, the variable insurance type has the least influence in determining the insurance claims amount.

Figure 4.22 is a variable importance plot for the Ghanaian dataset of the fitted RF model. The figure on the left shows the plot of the various attributes against their respective %IncMSE and the plot on the right shows plot the importance

University of Ghana http://ugspace.ug.edu.gh

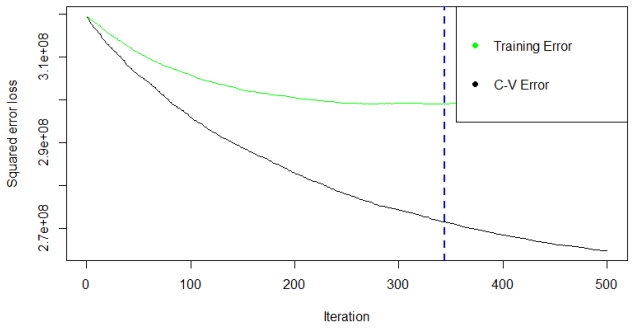


Figure 4.23: Plot of the Optimum Number of Iterations

of the various attributes against their respective %IncNodePurity.

4.5.3 Gradient Boosting

In fitting the GB model, we use the Gaussian loss function with 500 iterations. The result of this can be seen in figure 4.23. The 7 predictors utilized by the GB model all had a non-zero influence. But based on the cross-validation results, we discover that the best iteration on the training dataset is 344. The shrinkage value was set at 0.01, the interaction depth at 4, and the minimum number of observations was set at 10 for each terminal node.

Table 4.18 is a table showing the results of the relative influence of the variables used by the GB model.

Variable	Relative Inf.
Vehicle Age	29.46
Usage Type	22.09
Vehicle Make	21.49
Category	13.68
Purpose	7.19
Location	3.89
Insurance Type	2.19

Table 4.18: Relative Influence of the Variables

From the table, it can be seen that the variable Vehicle age, usage type, and vehicle make are the most important variables in growing the gradient boosted models with relative influence 29.46%, 22.09%, and 21.49% respectively. The

University of Ghana http://ugspace.ug.edu.gh

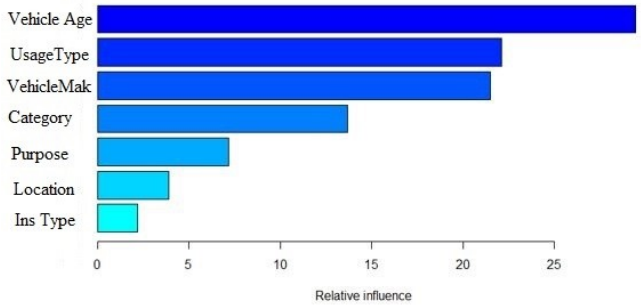


Figure 4.24: Relative Influence of the Variables

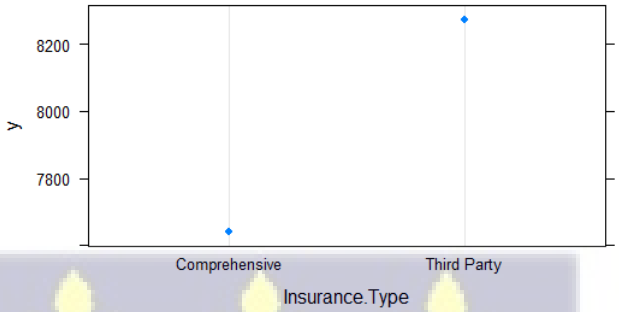


Figure 4.25: Partial Dependence Plot of the Variable Insurance Type

next important variables are the category, purpose, and location with relative influence 13.68%, 7.19%, and 3.89%. The variable with the least importance is the Insurance type with a relative influence of 2.19%. This is graphically represented in figure 4.24.

4.5.3.1 Partial Dependence Plots

From figure 4.25 the attribute that contributed the most in determining the claims amount is the Third Party under the variable Insurance type.

However, figure 4.26 also show the attributes Ambulance/Hearse, Corporate Buses, and Tankers contributed the most in that order in determining the claims amount under the variable usage type.

Figure 4.27 shows that the attribute corporate under purpose contributed the most in determining the claims amount. The next three attributes that con-

University of Ghana http://ugspace.ug.edu.gh

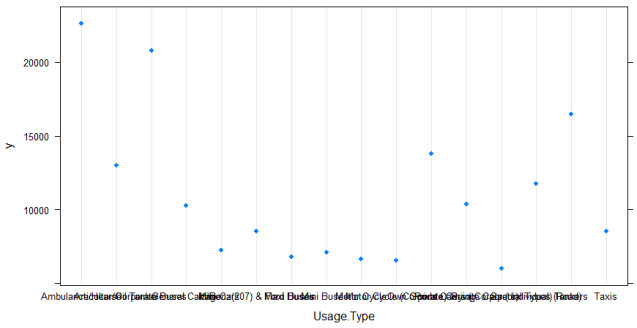


Figure 4.26: Partial dependence plot of the Variable Usage Type

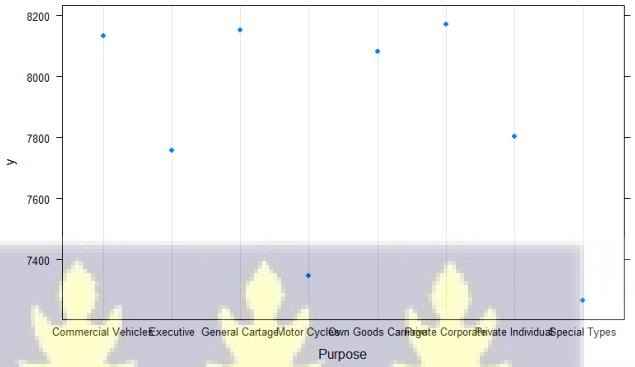


Figure 4.27: Partial Dependence Plot of the Variable Purpose

tributed the most are the general cartage, commercial vehicle, and own goods carriage respectively. The special type is the attribute that contributed the least.

Figure 4.28 shows that the vehicles with lower ages contributed the most in determining the insurance claims amount. While the vehicle with higher ages contributed the least in determining the claims amount.

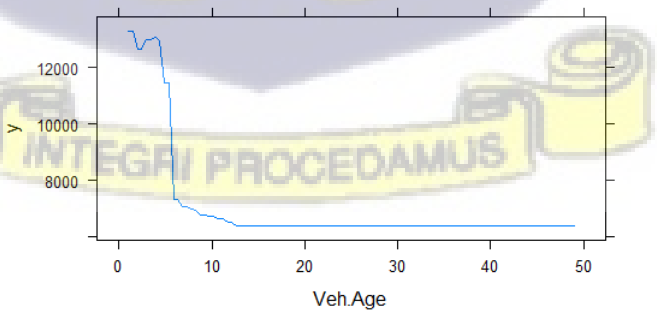


Figure 4.28: Partial Dependence Plot of the Variable Purpose

University of Ghana http://ugspace.ug.edu.gh

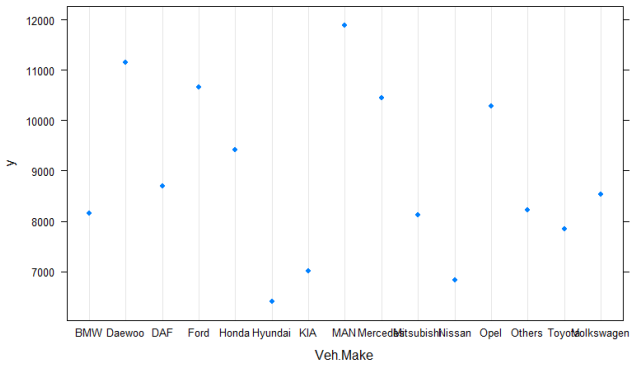


Figure 4.29: Partial Dependence Plot of the Variable Vehicle Make

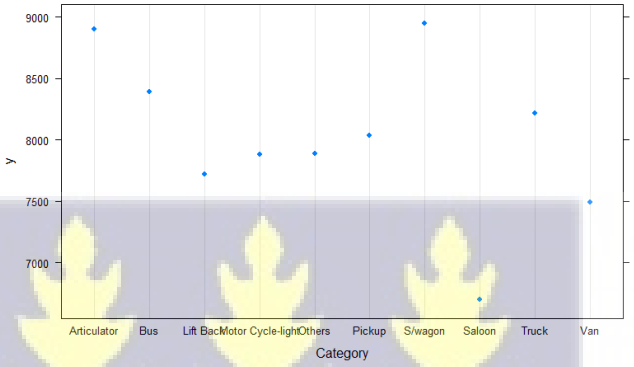


Figure 4.30: Partial Dependence Plot of the Variable Category

Figure 4.29 shows that the attribute MAN contributed the most in determining the claim amount. Other attributes like Daewoo, Ford, Mercedes, and Opel also contributed to determining the claim amount. The attribute that contributed the least under the variable vehicle make is Hyundai.

Figure 4.30 shows that Articulator and Station wagon were the attributes that contributed the most under the variable category in determining the claim amount. The attribute that contributed the least is Saloon.

Figure 4.31 shows that the attribute that contributed the most to determining the insurance claim amount under the variable location is the VR. On the other hand, the attribute that has the least contribution in determining insurance claims is AS.

University of Ghana http://ugspace.ug.edu.gh

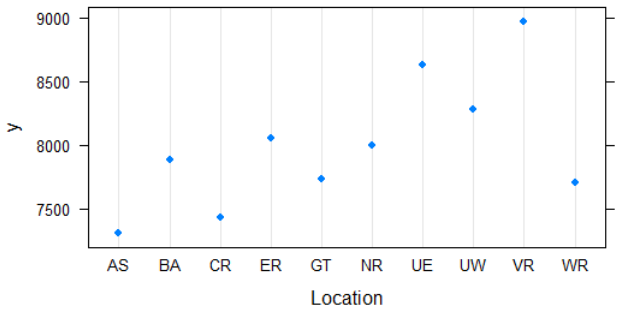


Figure 4.31: Partial Dependence Plot of the Variable Location

4.6 Performance Measures of the Machine Learning Models

In this section, the results of the performance measures obtained from the selected Tree-Based ML Models (CART, RF, and GB) for both the Nigerian and Ghanaian Data are discussed.

	Training Data			Testing Data		
Metrics	CART	RF	GBM	CART	RF	GBM
R2	**0.30**	0.18	0.16	0.10	**0.12**	0.08
MAE	**14028049**	16947025	14125562	17979177	**15491182**	17682446
MSE	**4.20e+15**	2.40e+16	2.65e+16	5.22e+16	4.47e+15	**2.10e+15**

Table 4.19: Performance Measures NG Data

From table 4.19, when the training data was used to test the predictive performance, the R-Squared reveals that the CART model performed better than RF and GB. The CART model had an R-Squared of 0.30 followed by the RF with an R-Squared of 0.18 and the GBM with an R-Squared of 0.16. However, for the testing data, the RF performed better than both the CART and GBM with an R-Squared of 0.12, while the CART and the GBM had an R-Squared of 0.10 and 0.08 respectively.

In measuring the predictive performance of the Models again, the MAE was used. For the training data, the CART had the lowest MAE of 14028049 and the RF had GBM had an MAE of 16947025 and 14125562 respectively. This demonstrates

University of Ghana http://ugspace.ug.edu.gh

that the CART model performed better on the training data than the RF and the GBM. Yet for the testing data, the MAE of the RF was the Lowest (15491182) followed by GBM (17979177) and the CART (17682446).

When the MSE was used to test the performance of the models, the CART came first above the RF and GBM based on the training dataset. For the testing data, the GBM has the lowest MSE followed by RF and CART respectively. On average anytime the training data was used the predictive accuracy of the CART model was higher based on the performance measures than the others. However when the testing data is used the predictive performance of the GBM and the RF was better than the CART.

	Training Data			**Testing Data**		
Metrics	CART	RF	GBM	CART	RF	GBM
R2	0.10	**0.45**	0.17	0.02	**0.16**	0.09
MAE	7698.93	**5963.51**	7492.59	8691.12	**7245.71**	8385.81
MSE	286560691	**232725012**	297934973	336314186	**190495194**	271538283

Table 4.20: Performance Measures GH Data

For the Ghanaian dataset, when the training data was used to test the predictive performance of the models, the RF model was the best based on the highest value of R-Squared of 0.45 percent. The GBM had an R-Squared of 0.17. However, the CART was the worst among the models with an R-Squared of 0.10 percent. Also, the RF model performed better on the testing data with an R-Squared of 0.16 followed by the GBM with 0.09. The CART had the least performance in terms of R-Squared with 0.02.

In measuring the predictive performance of the Models again, the MAE was used. For the training data, the RF model had the lowest MAE of 5963.51 and the CART and GBM had an MAE of 7698.93 and 7492.59 respectively. This demonstrates that the RF model performed better on the training data than the CART and the GBM. Yet for the testing data, the MAE of the RF was the Lowest (7245.71) followed by GBM (8385.81) and the CART (8691.12).

When the MSE was used to test the performance of the models on the Ghanaian

University of Ghana http://ugspace.ug.edu.gh

dataset, the RF came first above the CART and GBM based on the training dataset. For the testing data, the RF still had the lowest MSE followed by CART and GBM respectively. From the discussion above we can see that the RF model was the overall best model in all the performance measures both for the training and the testing datasets.

University of Ghana http://ugspace.ug.edu.gh

Chapter 5

Summary, Conclusion and Recommendation

5.1 Summary of Findings

This thesis analyzed two separate datasets of insurance claims. The Nigerian and the Ghanaian Dataset. The research compared the predictive performance of selected Tree-Based ML Methods (CART, RF & GBM) based on both training and testing datasets. Therefore, the results and findings are summarized in the following subsections.

5.1.1 The Nigerian Insurance Data

Based on the descriptive statistics, the researcher found out that the Insurance claims data is highly right skewed which conforms to most previous studies (Lee & Shi, 2019). The researcher also found out that very high claims occur, but not as frequent. Although, this was expected because the dataset contains the large claims paid for the period 2011 to 2018.

However, in fitting the Tree-Based ML models on the Nigerian dataset, this study found out that that the CART model did perform above the other models in terms of R-Squared based on the training sample. This can be attributed to the fact that when a model is exposed to a familiar data, the model's predictive accuracy may increase. But when the selected models were subjected to the unfamiliar data (i.e. the testing data) the RF performed better than the CART and the

University of Ghana http://ugspace.ug.edu.gh

GBM.

In terms of the Mean Absolute Error based on the training data, the CART also performed better that the RF and GBM. But for the testing data, this study found out that the RF model arose as the candidate model to predict insurance claims data based on the MAE. Another statistical metric used to analyze the performance of the models was the Mean Absolute Percentage Error. Using this metric to test the predictive accuracy of the selected models on the training data, we found out that the GBM was the candidate model but then, the CART model performed a little above the RF model. On the other hand, the RF model was the best performing above the CART and GBM model based on the testing data.

The other statistical metric used to assess the performance of the models was the Mean Squared Error. Based on this statistical measure of performance, we found out that, the CART model had the lowest MSE than the GBM and the RF model. This implies that in terms of the MSE, the CART model performed above the RF and GBM models.

It should be noted that the model that is able to predict and unseen data is considered a more powerful ML Algorithm, in our case therefore, using the testing data, we can see that the RF models is the best model in terms of the R-Squared and Mean Absolute Error. But the GBM arose as the model with a higher predictive performance in terms of the Mean Squared Error.

5.1.2 The Ghanaian Insurance Data

Similarly, the Ghanaian dataset was also highly right skewed based on the descriptive statistics given in chapter four. The same procedure was followed to fit the Ghanaian dataset based on the selected Tree-Based ML Algorithms. Therefore, this study found out that the RF model did performed significantly above the CART and GBM models in terms of R-Squared based on the training data.

University of Ghana http://ugspace.ug.edu.gh

Again the RF model did perform well in terms of the testing data and had the highest R-Squared than the CART and GBM model. For the testing data, the RF model had the lowest Mean Absolute Error when compared to the CART and GBM. Also when we used the testing data, the RF Model still was the best model that had the lowest Mean Absolute Error. In short, the RF is the overall best model using both the training and testing data in all other metrics used on the Ghanaian dataset.

5.2 Conclusion and Recommendation

Based on the findings of this study, we conclude that the RF model had the overall best performance based on the performance metrics. For the Nigerian insurance data, the RF model was the best model in terms of R-Squared and Mean Absolute Error, while the GBM is the most suitable model based on the Mean Squared Error . As for the Ghanaian dataset the RF model was the best in all the performance metrics. This can be attributed to the fact that the number of explanatory variables available for the Ghanaian dataset are a bit higher than that of its Nigerian counterpart. Previous researches confirm that the Tree-Based ML models (especially RF) can handle a very large volume of data especially when the number of explanatory variables greatly exceeds the number of observations (Dewi, Murfi & Abdullah; 2019). Although, in our case, the datasets available had only a few variables, yet our results show that the RF model had the overall best predictive accuracy when subjected to the testing dataset.

Furthermore, interest in the non-life insurance industry has been on the rise due to the large number of risks associated with the various types of insurance policies. As a result, insurers need to have a reliable approach/method to estimate insurance claims in order to cover for anticipated risks and contingencies, to re-

University of Ghana http://ugspace.ug.edu.gh

duce the likelihood of financial loss, to set adequate reserves, and ultimately to determine insurance premium according to the level of risks of policy holders. In the light of this, Tree-Based Machine Learning (ML) Algorithms are very powerful models used to analyze big data especially insurance data and actuaries use these models to predict insurance claims.

In conclusion, therefore, it is worthy to emphasize that the prediction of claims amount is one of the most important task for actuaries and insurance companies. This is because insurance entails managing a wide variety of risks. But in this analysis, the researcher faced numerous challenges in the process of obtaining the data and approached several insurance companies to obtain a richer dataset with more variables to conduct this analysis, but to no avail. This can be attributed to the fact that insurance companies (especially in Nigeria and Ghana) are very skeptical about releasing their insurance claims data possibly because of the fear of abuse of information and some other reasons. So therefore, this thesis analyzed the insurance claims data with the limited number explanatory variables available. This research validates the fact that Tree-Based ML models perform poorly, when explanatory variables in an insurance data are very few.

In the light of this, the study put forward the following recommendations:

- Insurance companies especially in Nigerian need to allow researchers have access to insurance claims data in order to help them conduct analysis so that researchers and actuaries can come up with a suitable model that can accurately predict and forecast insurance claims severity amount. This would significantly allow them determine an actuarially fair premium so as to stay competitive in the insurance industry.

- Where additional variables are available, it is also recommended that further research should be carried out on a bigger data having more explanatory

University of Ghana http://ugspace.ug.edu.gh

variables. And also the Tree-Based ML models should be compared with distributional ML models.

University of Ghana http://ugspace.ug.edu.gh

References

Achieng, O. M., & No, I. (2010). Actuarial modeling for insurance claim severity in motor comprehensive policy using industrial statistical distributions. In *International Congress of Actuaries, Cape Town* (Vol. 712).

Adesina, O. S., Adeleke, I., & Oladeji, T. F. (2016). Using Extreme Value Theory to Model Insurance Risk of Nigeria's Motor Industrial Class of Business. *The Journal of Risk Management and Insurance, 20*(1), 40-51.

Adesina, O. S. (2017). Modelling Claims of Nigeria Oil & Gas Insurance Class of Business with Extreme Value Theory. *Annals Computer Science Series* (Vol. XV)

Ademunigbohun S.S. and Oreshile S.A. (2014). Risk Attitude and Demand for Motor Insurance: An Examination of selected Motorist in Lagos State, Nigeria. *Developing Country Studies, 4*(2).

Adebayo, A. D., & Sunday, A. O. (2017). Modeling Claims of Nigeria Oil & Gas Insurance Class of Business With Extreme Value Theory. *Annals. Computer Science Series, 15*(2).

Adeleke, I. A., & Ibiwoye, A. (2011). Modeling claim sizes in personal line non-life insurance. *International Business & Economics Research Journal (IBER), 10*(2).

Anderson, W., Guikema, S., Zaitchik, B., & Pan, W. (2014). Methods for estimating population density in Data-limited areas: Evaluating regression and tree-based models in Peru. *PloS one, 9*(7).

Bartz-Beielstein, T., & Markon, S. (2004). Tuning search algorithms for real-

University of Ghana http://ugspace.ug.edu.gh

world applications: A regression tree based approach. In *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No. 04TH8753)* (Vol. 1, pp. 1111-1118). IEEE.

Balasooriya, U., & Low, C. K. (2008). Modeling insurance claims with extreme observations: transformed kernel density and generalized lambda distribution. *North American Actuarial Journal, 12*(2), 129-142.

Brazauskas, V., & Kleefeld, A. (2016). Modeling severity and measuring tail risk of Norwegian fire claims. *North American Actuarial Journal, 20*(1), 1-16.

Besler, E., Wang, Y. C., & Sahakian, A. V. (2019). Real-time Radio frequency ablation lesion depth estimation using multi-frequency impedance with a deep neural network and tree-based ensembles. *IEEE Transactions on Biomedical Engineering, 67*(7), 1890-1899.

Baumgartner, C., Gruber, L. F., & Czado, C. (2015). Bayesian total loss estimation using shared random effects. *Insurance: Mathematics and Economics, 62*, 194-201.

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

Bortoluzzo, A. B., Claro, D. P., Caetano, M. A. L., & 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*(1), 37-47.

Brockman, M. J., & Wright, T. S. (1992). Statistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries, 119*(3), 457-543.

Czado, C., Kastenmeier, R., Brechmann, E. C., & Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial*

University of Ghana http://ugspace.ug.edu.gh

Journal, 2012(4), 278-305.

Chukwudum, Q. C. (2019). Reinsurance Pricing of Large Motor Insurance Claims in Nigeria: An Extreme Value Analysis. *International Journal of Statistics and Probability, 8*(4).

Chen, X., Wang, M., & Zhang, H. (2011). The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 55-63.

David, M. (2015). Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance, 20*, 147-156.

Davoudi Kakhki, F., Freeman, S. A., & Mosher, G. A. (2018). Analyzing large workers' compensation claims using generalized linear models and Monte Carlo simulation. *Safety, 4*(4), 57.

Dewi, K.C., Murfi, H., & Abdullah, S. (2019). Analysis Accuracy of Random Forest model for Big Data- A Case Study of Claim Severity Prediction in Car Insurance. In *2019 5th International Conference on Science in Information Technology (ICSITech)* (pp. 60-65). IEEE.

Durica, M., Frnda, J., & Svabova, L. (2019). Decision tree based model of business failure prediction for Polish companies. *Oeconomia Copernicana, 10*(3), 453-469.

Escoto, B. (2013). Bayesian claim severity with mixed distributions. *Variance, 7*(2), 110-122.

Garrido, J., Genest, C., & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics, 70*, 205-215.

Goldburd, M., Khare, A., & Tevet, D. (2016). Generalized Linear Models for

University of Ghana http://ugspace.ug.edu.gh

Insurance Rating. *Casualty Actuarial Society, CAS Monographs Series, 5.*

Guelman, L., Guillén, M., & Perez-Marin, A. M. (2014). A Survey of Personalized treatment models for pricing strategies in insurance. *Insurance: Mathematics and Economics, 58*, 68-76.

Guo, L. (2003). Applying data mining techniques in property/casualty insurance. In *in CAS 2003 Winter Forum, Data Management, Quality, and Technology Call Papers and Ratemaking Discussion Papers, CAS.*

Henckaerts, R., Côté, M. P., Antonio, K., & Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal, 25*(2), 255-285.

Hong, L., Kuffner, T., & Martin, R. (2016). On prediction of future insurance claims when the model is uncertain. *Available at SSRN 2883574.*

Jain, N. Towards Machine Learning: Alternative Methods for Insurance Pricing–Poisson-Gamma GLM's, Tweedie GLM's and Artificial Neural Networks.

Klusowski, J. M. (2019). Analyzing CART. *arXiv preprint arXiv:1906.10086.*

Kuo, K., & Lupton, D. (2020). Towards Explainability of Machine Learning Models in Insurance Pricing. *arXiv preprint arXiv:2003.10674.*

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). *Loss models: from data to decisions* (Vol. 715). John Wiley & Sons.

Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association, 58*(302), 415-434.

University of Ghana http://ugspace.ug.edu.gh

Ng'elechei, J. K., Chelule, J. C., Orango, H. I., & Anapapa, A. O. (2020). Modeling Frequency and Severity of Insurance Claims in an Insurance Portfolio. *American Journal of Applied Mathematics and Statistics, 8*(3), 103-111.

Oyewole, O., Adesina, O. S., & Adekola, L. O. (2018). On Extreme Value Theory in Modeling Nigeria Marine and Aviation Insurance Class of Business. *AU eJournal of Interdisciplinary Research (ISSN: 2408-1906)*, 3(1).

Obiedat, R. Developing Decision Tree based Models in Combination with Filter Feature Selection Methods for Direct Marketing.

Olbricht, W. (2012). Tree-Based methods: a useful tool for life insurance. *European Actuarial Journal, 2*(1), 129-147.

Omerašević, A., & Selimović, J. (2020). Classification Ratemaking Using Decision Tree in the InsuranceMarket of Bosnia and Herzegovina. *The South East European Journal of Economics and Business, 15*(2), 124-139.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems, 9*(2), 181-199.

Rokach, L., & Maimon, O. Z. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

Ramon Callao, L. (2016). *Using tree-based models to predict drunk driving at Police preventive checkpoints* (Master's thesis, Universitat Politècnica de Catalunya).

Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance Claims. *Insurance: Mathematics and Economics, 64*, 417-428.

Schröder, W. (2006). GIS, geostatistics, metadata banking, and tree-based models for data analysis and mapping in environmental monitoring and epidemiology. *International Journal of Medical Microbiology, 296*, 23-36.

University of Ghana http://ugspace.ug.edu.gh

Soner, O., Akyuz, E., & Celik, M. (2018). Use of tree based methods in ship performance monitoring under operating conditions. *Ocean Engineering, 166*, 302-310.

Staudt, Y., & Wagner, J. (2021). Assessing the Performance of Random Forests for Modeling Claim Severity in Collision Car Insurance. *Risks, 9*(3), 53.

Su, X., & Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PloS one, 15*(8), e0238000.

Tober, S. (2020). Tree-based Machine Learning Models with Applications in Insurance Frequency Modelling.

Quan, Z. (2019). Insurance Analytics with Tree-Based Models.

Quan, Z., Wang, Z., Gan, G., & Valdez, E. A. (2020). Hybrid Tree-based Models for Insurance Claims. *arXiv preprint arXiv:2006.05617*.

Wüthrich, M., Non-Life Insurance: Mathematics & Statistics, March 14, 2017. *Available at SSRN: https://ssrn.com/abstract=2319328*

Wüthrich, M. V., & Christoph, B. (2019). Data Analytics for Non-Life Insurance Pricing (Swiss Finance Institute Research Paper No. 16-68). *URL https://ssrn. com/abstract, 2870308*.

Valecký, J. (2013). GLM analysis applied on claim severity of motor hull insurance portfolio: an empirical study. *Actuarial Science in Theory and in Practice*, 161.

Zounemat-Kermani, M., Stephan, D., Barjenbruch, M., & Hinkelmann, R. (2020). Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models. *Advanced Engineering Informatics*, 43, 101030.