KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

INSTITUTE OF DISTANCE LEARNING

Decision Tree as a Predictive Modeling Tool for General Insurance Claims

By

NIMO, NICHOLAS

Index Number

PG1477213

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF M.SC ACTUARIAL SCIENCE

March, 2016

Declaration

I hereby declare that this submission is my own work towards the award of the M.Sc 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 acknowledgment had been made in the text.

Nicholas Nimo	.....	.....
Student	Signature	Date

Certified by:

Nana Kena Frempong	.....	.....
Supervisor	Signature	Date

Certified by:

Prof. S.K Amponsah	.....	.....
Head of Department	Signature	Date

Dedication

I dedicate this work to my Mother Mrs. Rosina Antwiwaa , Lovely Son Joseph Kofi Korkor and His Mother Ernestina Sarfoa

Abstract

The study explored the relationship between the risk factors of policyholders of an insurance company and the occurrence of claim. These relationships were investigated by using decision trees developed by Breiman et al (1984) also referred to as CART analysis. The R software (Rattle) version 0.99.483 was used to analyze the data using the rpart component.The Analysis corroborated the view held by Yeo et al. (2001) and Mayhew et al. (2003) that a major risk factor that affects the occurrence of claim is the age of policyholders and also proved and ascertained Tranter and Warn (2008) views, that a group of young drivers have higher risk of accident than the same group of older drivers. Most importantly the analysis revealed that the prediction of whether a policyholder would make a claim in the coming year depends in part and in whole on the claim history of the policyholder.

Acknowledgments

I am most grateful to the Almighty God for His divine guidance and protection in the accomplishment of this thesis. I also wish to express my deepest gratitude and appreciation to my supervisor, Nana Kena Frempong for his mentorship, guidance and auspices even at tougher times throughout the period of completion of this thesis.

It is my pleasure to thank Dr. S.A Opoku, Prof. S.K. Amponsah, Prof. I.K. Dontwi, and all the lecturers of the Department of Mathematics who have in diverse ways supported me to accomplish this thesis

My sincere appreciation goes to Mr. Kobena Francis Addison, Cecil Ribeiro, Festus Awuah, and all staff of Quality Insurance Company, who have supported me in guidance and prayer.

To Mr. Benjamin Alfred Markin Yankah, Chief Actuary (NHIS), Mr. Mensah Bonsu Kakari (a private business man) and all my friends, I say may the good Lord bless you for your support, motivation and encouragement to see my dream come true.

Finally, thanks to my late father, Mr. Joseph Yaw Korkor and entire family members whose encouragement, guidance, affection and confidence in me has always given me strength and confidence in life to strive higher.

Contents

List of Abbreviation

AIC ..... Akaike Information Criteria

BIC ..... Bayesian Information Criteria

CRM ..... Classical Regression Model

CART ..... Classification and Regression Tree

CHAID ..... CHi-squared Automatic Interaction Detection

QUEST ..... Quick, Unbiased, Efficient Statistical Trees

NIC ..... National Insurance Commission

US AID ..... Automatic Interaction Detection

THAID ..... Theta Automatic Interaction Detection

List of Tables

List of Figures

Chapter 1

Introduction

Background of the Study

Risk classification in modern risk management helps in solving the problem of information asymmetry and moral hazards. The decisive theories of Rothschild & Stiglitz (1976) and Wilson (1977) predicts a positive association between the probability of a policyholder making claim and the munificence of his insurance contract. This reflects adverse selection between the insurer and the policyholder, "which leads to a sub-optimal allocation of risk within a risk class defined by characteristics observed by the insurer". Crocker & Snow (1986)

The risk exposure of the insurance industry in Ghana has increased tremendously NIC (2014), due to varied of factors which includes; high element of moral hazard, increase in the cost of claim, inflationary trend, pecuniary risk of the subject matter of insurance. The increasing risk that is being carried by insurers in Ghana facilitated the increment of motor insurance premium by Ghana Insurers Association to over 400% in 2015.

"General coverage that is given by the insurer decreases the expected rate of the occurrence of accident and therefore the incentives for safety" Shavell (1979); Holmstrom (1979). Meaning more insurance cover, which do not consider the risk at hand that may lead to occurrence of claim. It can be predicted that there is a positive correlation between risk and the extent of insurance cover within a risk class.

This thesis used decision tree, a form of machine learning to classify policyholders into groups based on their perculiar risk characteristics. In recent times there has been many innovative research and writings on the tree creation algorithms, to cite a few; Buckinx, Moons, Van Den Poel, & Wets (2004), Tan, Steinbach, & Kumar (2006), and Ngai, Xiu, & Chau (2009).

There have been various applications of this modeling approach in the industry such as; claim processing and management, detection of fraudulent claims, allocation of loss reserves, underwriting, and retail marketing campaigns. Predictive modeling goes to the next step and anticipates the future so that appropriate action can be taken and resources assigned earlier in the business process in order to try and achieve better outcomes.

Problem Statement

Insurance companies all over the world operate in a situation of uncertainty, thus lives in a myth of "the unknown future". In recent times there have been tremendous increase in the exposure of insurance companies in Ghana. These factors range from risk carried by the subject matter of insurance, the moral hazard of the insured and economic factors. Also due to societal advancement, claim knowledge has grown and every policyholder wants to derive the "main benefit" of insurance, "claim-making"; which has undoubtedly increased the number of claims and keep rising over time with consequential increase in insurance fraud.

The insurance company that bears the risk must have enough funds and reserves to pay claims; which indirectly means insured customers must pay equitable premium to commensurate the risk. There exits huge amount of data on risk factors and relationship between that and the claim must be gathered by use of data mining tools to be able to predict the behavior of an insured with

specific named characteristics.

The risk of frequent claim may vary or be the same (fundamental) across various groups of insurance coverage. These individual risk factors affect claims and are not normally accounted for at time of underwriting thereby affecting the premium-claim distributions of the insurer in the long run.

There is problem of developing a relationship between the occurrence of claim and the risk factors of a policyholder and/or the subject matter of the insurance. This thesis delves into developing this relationship, and associating probabilities to their predictive powers of occurrence of claim.

Objectives of the study

The main objectives of the study are as follows:

(i) To explore the prevailing factors underlying claim occurrence of the two major classes of policyholders; Individual and Corporate by use of the CART analysis;

(ii) Develop a predictive model for each of the classes in (i)given the risk characteristics by use of CART analysis;

(iii) Validate each of the model in (ii) by using an appropriate statistical package.

Methodology

Decision tree analysis is the major statistical tool that would be used to group policyholders into risk classes depending on their claim history spanning the period 2012-2014.

The main software for the analysis is the R Statistical software (version 0.99.483 – © 2009-2015), the 'Rattle' component, using the rpart algorithm; which checks all possible splitting variables, together with all possible values

to be used to split the nodes of the tree. In choosing the best splitter, the algorithm seeks to maximize the average "purity" of the two childnodes. Through a process of tree building and pruning of trees, an optimal tree would be obtained.

Justification of Study

Automobile insurance is the most common form of general insurance in Ghana with high frequency of claim arrival numbers. The first and utmost concern of any insurer when establishing premium is to ensure that the premium is sufficiently large to be able to fulfill her part of the contract when called upon. The insurer is not privy to the inherent characteristics or moral hazard of the prospective Policyholders prior to signing the contract. Overtime the characteristics of various policyholders unfold as they put in comprehensive or third party claim.

These inherent or latent characteristics which may be common among insured are the critical factors that help in the actuarial estimation of the relationship among these factors to predict claim counts of insured with some named characteristics. Moreover, the current demand for increase in premium by insurance due to frequent claim occurrence must bring our attention to critically classifying insured into groups to ensure adequate premium is paid to match the risk at hand. This means bonus scheme to award groups with good claim record, and penalty for those with worst record.

Limitation of Study

The major constraint on the thesis is availability of data. The nature of the research requires huge amount of claim data in order to be able to improve the predictive power of the model.

Chapter 2

Literature Review

Introduction

In this chapter, we review existing literature on general insurance and risk, as well as application of data mining and classification tree for predictive modeling.

Risk and Insurance

"Risk concerns the state of some financial relationship between an insurer and the insured, while uncertainty is simply a state of the mind". Roy and Roy (1994). Meldrum (2000), analyzed risk as “an unambiguous negative event that causes an actual loss or a reduction of the expected return”.

Insurance is a major tool for the transfer of risk. There is a wide range of views as to the way insurance operates. Mehr & Hedges (1963) defined insurance as “a device will be deemed to be insurance if it applies the law of large numbers so that the requirement for future funds to cover loss is predictable with reasonable accuracy and it provides some definite method for raising these funds by levies against the units covered by the scheme”.

Development of Insurance in Ghana

In 1924 Ghana began underwriting insurance products, with the establishment of the first insurance Royal Guardian Enterprise now known as the Enterprise Insurance Company Limited. The first indigenous private insurance company, the Gold Coast Insurance Company was established in 1955 to carry out the

business of insurance and later transformed to the State Insurance Company of Ghana (SIC) in 1962. In the year 1971, Eleven (11) more companies were established. After five years seven (7) more insurance companies, and one (1) reinsurance brokerage firms were established. Duodu, F.K & Amankwaa T. (2011). The National Insurance Commission, the sole institution that has been mandated to regulate and supervise insurance activities within the country was setup in the 1989 by the PNDC Law 229.

The major insurance companies in Ghana can be grouped into:

(i) Life Insurance

(ii) Non-Life Insurance, and

(iii) Composite Insurance (a combination of Life and Non-Life insurance).

According to the NIC annual report 2014 there: 26 Non-Life companies, 20 Life companies, 3 Reinsurance companies with over 40 Broking companies, one reinsurance broking company, one loss adjusting company and about 4,500 insurance agents.NIC (2014)
Insurance companies in Ghana offers various range of insurances such as; Automobile Insurance, Household Insurance, Fire & Allied Perils Insurance, Various Bond Insurances, Personal Accident, Marine etc. However, this research concentrates on the Automobile insurance which offers two-thirds of insurance companies' revenue in Ghana. NIC (2014)

Automobile insurance is defined as the type of insurance that is purchased for vehicles , trucks, motorcycles, and other road vehicles. The primary use of insurance is to provide financial protection against physical damage and/or bodily injury resulting from traffic collisions and against liability that could also result. Insurance is nowadays extended to cover loss as a result of other perils other than those discussed above.

In recent times, there has been an increased need for automobile insurance due to advancement in the legal regime. For example, in Ghana, the (Third Party Insurance) Act, 1958 (Act 42) makes it a crime to drive a motor vehicle on a public road without insurance covering third party liabilities. This demand for auto insurances has also increase the liabilities assumed by the insurance companies.

The main categories of Automobile Insurance are;

- Comprehensive insurance;
- Third Party Fire only
- Theft and third Party

These policies are further classified into Private or Commercial depending of the use of vehicle, either use for commercial activities or for non-commercial or self gratification. The laws in Ghana requires drivers to carry third party insurances; which covers the drivers liability towards third parties. However, comprehensive insurance cover which covers an own damage to your vehicle as a result of collision, accidental damage is optional if you own the vehicle outright.

Players in the Insurance Industry

The insurance industry like any other industry depends on the expertise of many individuals in their field, some of them are discussed as follows:

(i) Underwriters

Underwriting is the practice of assessing the eligibility of a customer in receiving coverage or protection for the risk of future loss that they born. An insurance underwriter evaluates the risk that is attached to a policy proposer. In practice, the underwriter determines if the particular insurance package would be of mutual benefit to the proposer and the insurance

company. The role of an underwriter requires knowledge in risk assessment, quantitative skills, a good communication skills , experience in research, together with computer skills.

(ii) Claims

Claims is a formal request made to an insurance company in returns for payment based on the extent of damage and the terms and condition of an insurance policy. The claims department of every insurance company has been said to be the "window to the insurance company". Therefore, there is the need to ensure that the section is properly managed.

Reported Insurance claims are carefully evaluated and given greatest attention, so as to examine the policy,elicit relevant information from the claim, match responses from claimant with that of eye witnesses, third parties and the police in order to establish liability. In arriving at an accurate decision, claims staff works with investigators and loss adjusters before making recommendations for payment.

(iii) Loss Adjuster

This is a claims specialist who investigates complex claim on behalf of insurance companies and any person that holds an insurance policy. The loss adjuster works on a damaged property claims to determine the property written off on insurance, review policies, get supporting documents to validate claims, and investigate site of the damage to determine if property damaged deserves compensation. They can be independent contractors or employees of an insurance firm.

(iv) Insurance Agent

An insurance agent is a sale represent person who act for and on behalf of an insurance company. They play a fundamental role in helping clients choose insurance policies that suit their needs and aide in the completion of a policy application. Payment of commission is their main source of remuneration,

since they are not permanent employees of any insurance company. They either works as freelancers to offer multi insurance packages from different insurance companies, or just get affiliation with one company.

(v) Insurance Broker

An intermediary between an insurer and insure is called insurance broker. They use their in-depth knowledge in risk and insurance to education the "proposer" appreciate the kind of insurance being purchased, the pricing aspects together with knowledge at time of claim.

(vi) Reinsurer

The reinsurer is a specialist company that accepts a risk portfolio from the direct insurer and shares in the premium and claim in a pre-determined manner. In short the reinsurer works to insure the insurer's liabilities in excess of its capacity. The reinsurer needs to have an in-depth knowledge in the insurance market and its relevant legalities to stand such responsibility. The financial capacity to pay claims, handle disputed claim and advice on various insurance issues are some attributes of the reinsurance company.

(vii) Risk Manager

Risk manager is a person who manages, assess and quantify business risk, taking into accounts measures to reduce or mitigate them. In insurance, the risk manager evaluate risks which endangers the company's funds or capacity to pay claims. Their work is done together with underwriters by assess the degree of risk the company can cover known as 'the risk appetite' of the company.

(viii) Actuary

A business professional that deals with the financial impact of a risk and uncertainty with mathematical probabilities, and thus predict future events. This is done with statistical data, demographics, financials, economic and social data in assessing risks related to financial planning. The

actuary should be able to communicate technical concept to non-technical individuals.

Claim Modeling

Yeo et al.(2001) in their study found that frequency and severity of claims should be modeled separately. They argued that insured groups significantly differ from one another by claim frequency and by claim severity(average claim cost). They ascertained that "even though female drivers tend to have more vehicular accidents, than males, these are usually low cost for insurance company; whilst the damage that is caused by a male driver is more likely to be an expensive claim.

In their paper, Murphy et al.(2000) used GLM for estimating the risk premium and elasticity of customer value models. Moreover, several studies have each identified that there is a core of risk factors that are equally important, regardless of country or the insurance company. Those risk factors are categorized into two major groups: motor vehicle's characteristics, and driver's characteristics. Yeo et al. (2001), Murphy et al.(2000), Tryfos (1980) and Wenzel & Ross (2005).

The risk factors are very essential in differentiating between profitable and non-profitable policyholders, therefore it would be more profitable to the insurance company to separate high risk drivers and low risk drivers, so that "every insured contributes his/her fair share towards the risk involved" Randal et al. (1978)., where premiums charged to policyholders reflect their expected losses or gains.

Many previous studies have concluded that Age and gender of drivers are statistically significant predictors of claim cost Yeo et al.(2001), Mayhew et al. (2003). Tranter & Warn (2008) in their study ascertained that a group of young

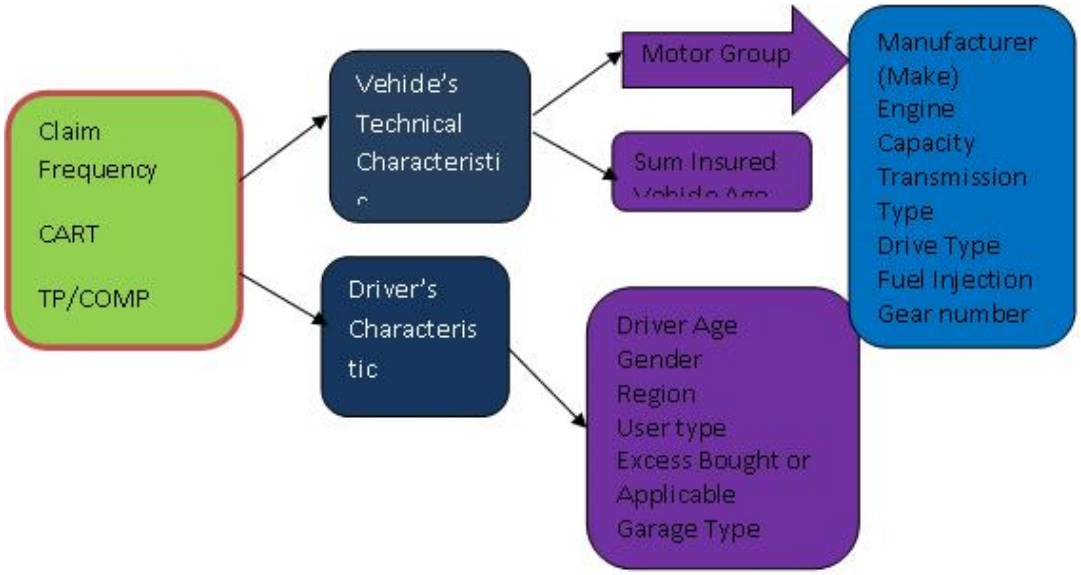

Figure 2.1: Risk Factors

men are hazardous drivers compared to the same size group of older drivers. The study affirmed that " youngsters tend to have more vehicle accidents which result in higher losses for insurance company". They gave two major reasons for such a pattern; firstly, young people lack driving skills and experience. Secondly, young drivers (both male and female) often get involve in car racing and other adventurous and risk - attracting events, as opposed to mature drivers, and therefore, their vehicle accidents record is greater.

Comparison of Data Mining Tools

Ahmed (2004); Carrier & Povel (2003); Mitra et al. (2002); Shaw et al. (2001) all described the various types of data mining modeling tools in predictive modeling. According to Xhemali, D., Hinde, C. J., & Stone, R. G. (2009). "Although some of the results of Naïve Bayes Classifiers are closed to the results from the Decision Trees classifier, Naïve Bayes Classifiers should not be considered superior to more complex techniques such as Decision Trees". According to the

paper, Decision trees may be computationally expensive for certain domains, however, they make up for it by offering a genuine simplicity of interpreting models, and helping to consider the most important factors in a dataset first by placing them at the top of the tree.

The neural network and other data mining tools are not so easy to understand from the visual representation. It is very difficult to create computer systems from them, and almost impossible to create an explanation from the model. Moreover, according to Moore, A. W. (2001), decision trees are the single most popular tool for data mining due to the following strengths:

- Decision trees are able to generate understandable rules.
- Decision trees perform classification without requiring much computation.
- Decision trees are able to handle both continuous and categorical variables.
- Decision trees provide a clear indication of which fields are most important for prediction or classification.

The Decision Tree

The Wikipedia defined a decision tree as a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility".
Gordon V. Kass (1980) used CHAID a type of decision tree technique, based upon adjusted significance testing. CHAID stands for Chi-squared Automatic Interaction Detection, based upon a formal extension of the US AID (Automatic Interaction Detection) and THAID (Theta Automatic Interaction Detection) procedures of the 1960s and 1970s, which in turn were extensions of earlier research, including that performed in the UK in the 1950s.

According to research in cognitive psychology, Kahneman, Slovic & Tversky (1982) observed “the ability to theoretically grasp and manipulate multiple chunks of information is limited by the physical and cognitive processing limitations of the short term memory portion of the brain”. With this, we need a high utilization of dimensional exploitation and presentation methods capable of preserving and reflecting high-dimensionality relationships in a readily comprehensible form so that the relationships can be easily appreciated, comprehend and applied by end users.

Berry & Linoff (1997); Chen, Hsu, & Chou (2009); in their respective writings found out that “for describing sequence of interrelated decisions or predicting future data trends, Decision Tree Model is used”.
In order to partition up a large collection of records into successively smaller sets of records by applying a sequence, Lee & Siau (2001) or as Berry & Linoff (2004) suggested the use the of simple decision rules. According to Ngai, Xiu, & Chau (2009), “decision tree technique has been proven to be among the top three popular technique of data mining in CRM is used”. The technique is capable of partitioning specific entities into classes based on feature of entities, Buckinx, Moons, Van Den Poel, & Wets(2004). In their paper, Choi, Ahn & Kim (2005) applied decision analysis method, "the Analytic Hierarchy Process (AHP)" to aggregate the opinions of the group decision makers on what are the relevant criteria for evaluating business values of rules and relative importance of those criteria. Tan, Steinbach, & Kumar (2006) described each tree to consists of three types of nodes. R. Sipulskyte (2012), used various methods for classifying policyholders based on their risk characteristics. He classified risk by using the following variables; The sum insured, location, driver’s age, gender, vehicle age, use of vehicle, garage type, voluntary excess amount, no claims discount, and time/inflation. Also, D. Pozzolo, A. et al. (2010) used classification tree for predicting bodily injury claim cost based on the characteristics of the insured

vehicle.

Many individuals have come up with innovative tree creation algorithms. Important ones come from Morgan & Sonquist (1963), Breiman et al. (1984), and Quinlan et al. (2008).

Nagy et al. (2010) in their paper "Tree-Based Methods as an Alternative to Logistic Regression in Revealing Risk Factors of Crib-biting in Horses" tried to establish the difference between tree based methods and logistic regression. Therneau T., Atkinson E.(1997) used the rpart component of R to carryout classification. In addition, Kajungu et al. (2012) used classification tree modeling to investigate drug prescription practices at health facilities in rural Tanzania. The aim of their study was to understand the factors influencing prescription patterns and to develop strategies to mitigate the negative consequences associated with poor practices in both the public and private sectors.

In their literature, Wei-Yin Loh & Yu-Shan Shih (1997) applied the use of exhaustive search method and QUEST for selecting variables that afford more splits. It was realized that;

1. The exhaustive search method is faster than QUEST when the number of classes $J = 2$. This is due to a short-cut algorithm that reduces the number of splits searched from $(2M - 1 - 1)$ to M Breiman et al. (1984), Theorem 4.5). This short-cut is only applicable when $J = 2$.

2. For $J > 2$ and $M > 4$, QUEST is faster than exhaustive search, with relative speed increasing exponentially with M and linearly with K.

3. The relative speed does not vary much with J for $J > 3$.

Austin (2008) studied classification trees grown using R and those grown using S-PLUS. i.e. R and S-PLUS are two statistical programming languages that share a similar syntax and functionality.

Ritschard G. (2006)., considered the deviance as a goodness-of-fit statistic that attempts to measure how well the tree is at reproducing the conditional distribution of the response variable for each possible profile, rather than the individual response value for each case.

Chapter 3

Methodology

Introduction

With the advancement in the use of computer technology especially in the field of statistical inference, one can easily make analysis of data with little knowledge about the statistical and mathematical concepts that underline it. This chapter takes us through an important exercise to acquire the knowledge and understanding of the theory and conceptual framework of the statistical method that is used to analyze the data efficiently and effectively.

Data Description

Longitudinal paid claims from 2010 to 2014 was collected from Quality Insurance Company Ltd. Microsoft Excel was used as the main software to organize the data into a tabular form as shown.

Table 3.1: Data Snapshot

CLAIM?	MAKE	USE	GENDER	VEH.AGE	PH.AGE
Yes	Nissan	PRIV	F	7	26
Yes	Renault	COMM.	M	6	34
No	Mitsubishi	PRIV.	M	7	35
Yes	Nissan	COMM.	M	11	36
No	Toyota	PRIV.	F	8	25
No	Mitsubishi	PRIV.	M	9	39
No	Renualt	PRIV.	F	4	32
Yes	Toyota	COMM.	M	6	42

For the purpose of predictive modeling into the coming year, 2015, the current claim record of the policyholders as at 2014 was used together with the corresponding characteristics of the insured or subject matter of insurance.

Decision Trees

The classification and regression tree (CART) algorithm was used by Breiman L. and Stone (1984) to fit trees, branches and leaves to data so as to observe predictive behavior of factors or variables under consideration.

The main output structure that evolve as a direct consequent of asking an ordered flow of questions is the Decision tree. The type of question that is asked at each step in the flow depends upon the answers to the previous questions of the sequence.

Node Impurity Functions

All of the allowable ways of splitting at each stage of continuous partitioning into subset of L are considered at each node of the tree. The split which would lead to the greatest increase in node "purity" is then chosen for the split. This can be achieved using the "impurity function"; the functions of the proportions of the learning sample belonging to the possible classes of the response variable. We choose the best split over all the variables with the aim to have as little purity as possible.

Accordingly, the best split is the one that reduces the node impurity the most.

Let $(P_1, P_2, \cdots, P_k)$ be $k >= 2$ classes, then, at any given node, ρ the impurity function at each node $imp(\rho)$ is defined as;

$$imp(\rho) = (\psi(1|\rho), \Psi(2|\rho), \cdots, \psi(k|\rho)) \tag{3.1}$$

where $P(k|\rho)$ is an estimate of $P(X \in \pi_k)$, i.e. the conditional probability that an observation X is in π_k given that it falls into node ρ.

It is required for $imp(\rho)$ to be a symmetric functions defined on the set of all K-tupples of probabilities $(p_1, p_2, \cdots, p_k)$ with unit sum, minimized at the points

$(1,0,0,...0)$; $(0,1,000,\cdots,0)$; ...; $(0,0,0,0,0,\cdots,1)$ and minimized at the point P= $(1/k, 1/k, ..., 1/k)$.

Now if $k = 2$, these conditions reduces to a symmetric $\psi(p)$, maximized at the point $p = 1/2$ with $\psi(0) = \psi(1) = 0$..

Various Impurity Measure Functions

1. **Resubsitution Error**

   The most obvious choice of impurity measure is the so-called resubstitution error. It measures what fraction of the cases in a node is classified incorrectly, if we assign every case to the majority class in that node. That is;

   $$imp(\rho) = 1 - max\ jp((j|\rho)) \tag{3.2}$$

   where $p(j/\rho)$ is the relative frequency of class j in node ρ. For a two-class problems we denote the classes by 0 and 1; P(0) denotes the relative frequency of class 0 and P (1) must be equal to 1-P(0) since the relative frequencies must sum to 1.

2. **The entropy function**:

   This is given by;

   $$imp(\rho) = -\sum_{i=1}^{k} P(i/\rho) log P(i/\rho)) \tag{3.3}$$

   Now if k=2, it implies that,

   $$imp(\rho) = -\sum_{i=1}^{2} P(i/\rho) log(P(i/\rho) \tag{3.4}$$

   $$= -\sum_{i=1}^{2} P(i/\rho) log(P(i/\rho)) \tag{3.5}$$

Now Let $P(1/\rho)$ =P, this implies $P(2/\rho)$ = 1-P

Therefore equation 3.5 becomes,

$$= -PlogP - (1-P)log(1-P) \tag{3.6}$$

3. **The Geni Index Function**

   This is given by; for all $(i \neq k)$,the Geni Index at any node ρ is measured by;

$$imp(\rho) = 1 - \left[sum_{i=1}^{k} P(i/\rho)\right]^2 \tag{3.7}$$

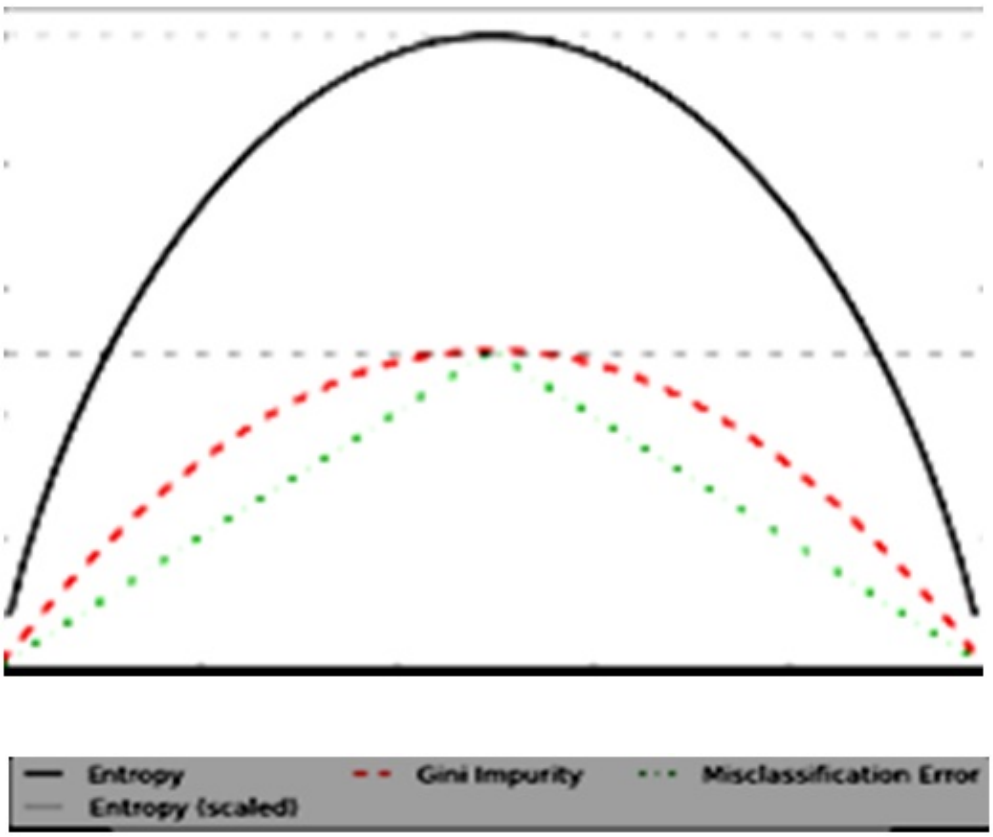

Figure 3.1: Chart of Impurity Measures

As shown Fig 3.1 above, all the 3 functions for measuring impurity at the node are concave, having minimum point at $p = 0$ and $p = 1$ and a maximum at $p = 0.5$.

However the following should be noted:

- The Gini Index is more likely to partition the data so that there is one relatively homogeneous node having relatively few cases.
- The Entropy tends to partition the data so that all of the nodes at any given split are about equal in size and homogeneity.

But practically, there is not much difference between these two types of measures of node impurity.

The Sets of Split Considered

Having looked at different criteria for assessing the quality of a split, we look at which splits are considered in the first place. We denote the set of explanatory variables (features) by $(x_1, \cdots x_p)$.

Variables may be numeric (ordered) or categorical. The set of splits that is considered are defined as follows:

1. Each split depends on the value of only a single variable;
2. If x is numeric, we consider all splits of type $x \leq a$ for all a ranging over$(-\infty, \infty)$.
3. If x is categorical, taking values in $V(x) = b_1, b_2, \cdots, b_L$, we consider all splits of type $X \in S$, where S is any non-empty proper subset of $V(x)$.

Splits on Numeric Variables

We can easily see there are only a finite number of distinct splits of the data. Let n denote the number of examples in the training sample. Then, if x is ordered, the observations in the training sample contain at most n distinct values $(x_1, x_2, \cdots, x_n$ of X. This means there are at most $n-1$ distinct splits of type $(x \leq a_m)$, $\quad m = 1, \cdots, n* \leq n$, where the a_m are taken halfway between consecutive distinct values of X.

Splits on Categorical Variables

For a categorical variable X with L distinct values there are $(2^{L-1}-1)$ possible splits to consider. Also, note that there are 2^{L-2} non-empty proper subsets of $V(x)$ (i.e. the empty subset and $V(x)$ itself are no splits at all). But the splits $(X \in S)$ and $(X \in S_c)$, where S_c is the complement of S with respect to $V(x)$, are clearly the same, so the number of different splits is only $\frac{1}{2}(2^{L-2})) = (2^{-1} \times 2^L) - 1) = (2^{L-1}-1)$

Choosing the Best Possible Split

Suppose at Node ρ, we apply split S so that proportions P_L of the observations drops down to the left daughter-node ρ_L and the remaining proportion P_R drops down to the right daughter-node ρ_R.

For example, suppose we have a data set 12 in which the response variable Y has two possible outcomes, “yes” and “no". Now assume that one of the possible splits of the input variable X_j is $X_j \leq a$ and $X_j > a$), where a is some value of X_j .

Then, we represent the 2x2 table which represent the number of responses in each Boolean (yes or no) as shown in table 3.5 with estimate for $P_L = \frac{N*1}{N**}$ and that of $P_R = \frac{N*2}{N**}$

Target	Yes	No	Row Total
Xj<=a	N11	N12	N1*
Xj>a	N21	N22	N2*
Column Total	N*1	N*2	N**

Now, the Entropy i(ρ) is given by the impurity measure;

$$imp(\rho) = -\left(\frac{N*1}{N**}\right)\log\left(\frac{N*1}{N**}\right) - \left(\frac{N*2}{N**}\right)\log\left(\frac{N*2}{N**}\right) \quad \text{i.e from (3.4)}$$

$$\text{Given that } x_j \leq a, \quad P_L = \frac{N11}{N1*} \quad \text{and} \quad P_R = \frac{N12}{N1*}$$

$$\text{Also for } x_j > a \quad P_L = \frac{N21}{N2*} \quad \text{and} \quad P_R = \frac{N22}{N2*}$$

Now the entropy estimate for the two daughter nodes as estimated as follows;

$$imp(\rho_L) = -\left(\frac{N11}{N1*}\right)\log\left(\frac{N11}{N1*}\right) - \left(\frac{N12}{N1*}\right)\log\left(\frac{N12}{N1*}\right) \tag{3.8}$$

$$imp(\rho_R) = -\left(\frac{N21}{N2*}\right)\log\left(\frac{N21}{N2*}\right) - \left(\frac{N22}{N1*}\right)\log\left(\frac{N12}{N1*}\right) \tag{3.9}$$

The goodness of split G_s at node ρ is given by the reduction in impurity gained by splitting the parent node into its daughter nodes, ρ_L and ρ_R i.e.

$$\triangle(Gs, \rho) = imp(\rho) - \frac{N1*}{N**} * imp(\rho_L) - \frac{N2*}{N**} * imp(\rho_R)$$

The variable X_j that gives the best split is the one that has the largest value of the above named equation.

After splitting the root (parent) node, we continue to divide the two daughter nodes using the same principle as illustrated above but with fewer variables than before. For instance, to further divide ρ_L or ρ_R , the partitioning process is repeated with a minor adjustment in order to arrive at an overall efficient split at any given node ρ.

The Basic Tree Construction Algorithm

This is to take an overview of the basic tree construction algorithm. The algorithm maintains a list of nodes (i.e. a set of examples) to be considered for expansion. The set of training examples are placed on this list of nodes. A current node is then selected from the list. Now for nodes which contain examples different from the classes (i.e. its impurity is larger than zero), then we find the best split and apply this split to the node. The resulting child nodes

are added to the list. The algorithm is finally halted when there are no further nodes to be expanded.

1) Nodelist ← training sample
2) Repeat (1)
3) current node ← select node from nodelist
4) nodelist ← nodelist-current node
5) If impurity (current node) >0 Then,
6) S ← candidate splits in current node
7) s* ← argmaxs ∈ S impurity reduction (s,current node)
8) child nodes ← apply(s*,current node)
9 nodelist ← nodelist ∪ child nodes
10) Until nodelist =

Box 3.1: Tree Construction Algorithm

Overfitting and Pruning

Once we know which applicants have defaulted and which have not, we can construe some complicated model that gives a perfect explanation.
By fitting the model perfectly to the data, we have "overfitted" the model to the data, and have in fact been modeling noise.
To avoid over fitting when we construct a classification tree, the following two approaches have to be implemented:

- Stopping Rules: Don't expand a node if the impurity reduction of the best split is below some threshold.

  The disadvantage of a stopping rule is that sometimes you first have to make a weak split to be able to follow up with a good split.

- Pruning: Grow a very large tree and merge back nodes.

Cost Complexity Pruning

Having built a large tree we then look at different pruned subtrees of this larger tree and compare their performance on a test sample.

The pruning of a tree T at a node ρ logical implies that ρ becomes a leaf node and all descendants of ρ are removed.

See Figure 3.4 and Fig 3.5 for the tree that results from pruning the major tree in figure 3.3 in node ρ_2. The branch T_{ρ_2} consist of node ρ_2 and all its descendants. The tree obtained by pruning T in ρ_2 is denoted by $T - T_{\rho_2}$.

A pruned subtree of T is any tree that can be obtained by pruning T in 0 or more nodes. If $T*$ is a pruned subtree of T, we write this as $T_0 \leq T$ or alternatively $T \geq T*$.

Now the number of pruned subtrees may become very large and it may not be feasible to compare them all on a test set.

The basic idea of cost-complexity pruning is not to consider all pruned subtrees, but only those that are the "best of their kind" as illustrated by the figures below.

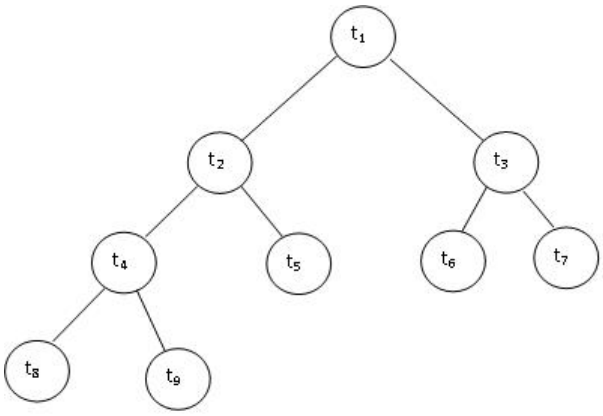

Figure 3.2: Tree T with leaf nodes $\hat{T} = \{t_5, t_6, t_7, t_8, t_9\}$, $|\hat{T}| = 5$

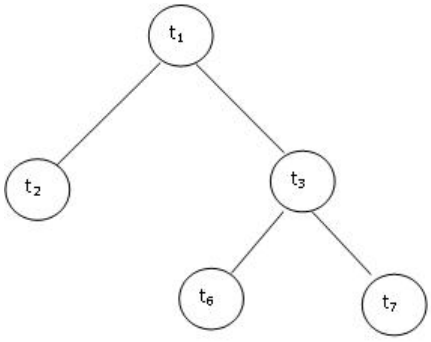

Figure 3.3: Branch $T_{\rho_2} : |\hat{T}| = \{t_2, t_6, t_7\}, |T_{\rho_2}| = 3$

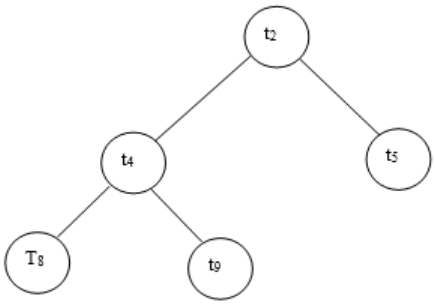

Figure 3.4: Branch $T_{\rho_2} : |\hat{T}| = \{t_5, t_8, t_9\}, |T_{\rho_2}| = 3$

Let R(T) denote the fraction of cases in the training sample that are misclassified by T. R(T) is called the re-substitution error of T).

Define the total cost $C\alpha(T)$ of tree T as

$$C\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{3.10}$$

Where R(T) is the re-substitution error, and $\alpha T*$ is the penalty for the complexity of the tree which denotes the set of leaf nodes of T, with α being the parameter that determines the complexity penalty.

When the number of leaf nodes increases by one i.e. one additional split in a binary tree), then the total cost (if R remains equal), increases with α.
Now Depending on the value of $\alpha \geq 0$, a complex tree that makes no errors may now have a higher total cost than a small tree that makes a number of errors.

Let T_{max} represent the large tree that is to be pruned to the right; Then given a fix value of α $\exists$ the smallest minimizing subtree $T(\alpha)$ of T_{max} satisfying the following conditions:

- $C\alpha(T(\alpha)) = \min\ T_{max}\ C\ \alpha(T)$
- If $C\alpha(T) = C\alpha(T(\alpha))$

then $T(\alpha) \geq T$

NB: The first condition says there is no subtree of T_{max} with lower cost than $T(\alpha)$, at this value of α.

The second condition says that if there is a tie, i.e. there is more than one tree that achieves this minimum, then we pick the smallest tree (i.e. the one that is a subtree of all others that achieve the minimum).

It can be shown that for every value of α there is such a smallest minimizing subtree. This implies that it cannot occur that we have two trees that achieve the minimum, but are incomparable.
Although α goes through a continuum of values, there is only a finite number of subtrees of T_{max}.

We can construct a decreasing sequence of subtrees of $T_{max} > T_1 > T_2 > T_3 > ... > t_1$ (where t_1 is the root node of the tree) such that T_k is the smallest minimizing subtree ($\forall\ \alpha\ \in [\alpha_k, \alpha_{k+1})$. This is an important result, because it means we can obtain the next tree in the sequence by pruning the current one. This allows the specification of an efficient algorithm to find the smallest minimizing subtrees at different values of α.

The first tree in the sequence, T_1 is the smallest subtree of T_{max} with the same restitution error as T_{max} (i.e. $(T_1 = T \text{ for } \alpha = 0)$
We prune (T_1) in these nodes to obtain T_2, the next tree in the sequence. Then we repeat the same process for this pruned tree, and so on until we reach the root node.

Choosing the Best Subtree

The various subtrees produced by the pruning algorithm serves as the set of subtrees required to model the data in order to obtain the classification tree.
The other aspect is the selection of the one which will hopefully have the smallest misclassification rate for future observations. Breiman et al. (1984) ordered two estimation methods, which is the independent test sample or cross-validation.

1. **Independent Test Set**

   This is used to estimate the error rates of the various trees in the nested sequence of subtrees, and the tree with minimum estimated misclassification rate can be selected to be used as the tree-structured classification model.

   For this purpose, the observations in the learning dataset (D) are randomly assigned to two disjoint datasets, a training dataset (L) and a test set (T), where $L \cap T = \Phi$ and $L \cup T = D$. Suppose there are n_T observations in the test set and that they are drawn independently from the same underlying distributions as the observations in L. Then the tree T_{max} is grown from the learning set only, and it is pruned from bottom up to give the sequence of subtrees $T_1 > T_2 > T_3 >, ..., > T_M$, and a class is assigned to each terminal node.

   Once a sequence of subtrees has been produced, each of the n test-set observations is dropped down the tree T_k. Each observation in T is then classified into one of the different classes. Because the true class of each observation in T is known, $R(T_k)$ is estimated by $R^{ts}(T_k)$, with $\alpha = 0$; that is $R^{ts}(T_k) = R^{re}(T_k)$, the resubstitution estimate computed using the independent test set. When the costs of misclassification are identical for each class, $R^{ts}(T_k)$ is the proportion of all test set observations that are

misclassified by T_k. These estimates are then used to select the best pruned subtree T_α by the rule, $R^{ts}(T\alpha) = \min\ R^{ts}(T_k)$ and $R(T\alpha)$ is its estimated misclassification rate.The standard error of $R^{ts}(T)$ is estimated as follows. When test set observations are dropped down the tree T, the chance that any one of these observations are misclassified is $p* = R(T)$. Thus, it is a binomial sampling situation with n_T Bernoulli trials and probability of success $p*$.

If $p = R^{ts}(T)$ is the proportion of misclassified observations in T, then p is unbiased for $p*$ and the variance of p is;

$$\frac{p*(1-P*)}{n_T} \tag{3.11}$$

And the standard error of $R^{rs}(T)$ is;

$$\left\{\frac{R^{ts}\,(1-R^{ts})}{n_T}\right\}^{\frac{1}{2}} \tag{3.12}$$

2. **Cross Validation**

For a V-fold cross-validation (CV/V), the learning dataset D is divided into V roughly equal-sized, disjoint subsets,$D = \cup D_v = 1$

Where $D_v \cap Dv* = \Phi, \quad v \neq v*$, and V is taken to be 5 or 10. Next, V different datasets are obtained from the D_v by taking $L_v = D - D_v$ as the vth training set and $T_v = D_v$ as the vth test set,$v = 1, 2, 3, ..., V$. The v-th tree $T_{max}(v)$ is grown using vth training set $L_v, v = 1, 2, 3, ..., V$. The value of the complexity parameter is fixed to a certain value. Let, $T(v)(\alpha)$ be the best pruned subtree of $T_{max}(v)$. Now, each observation in the v-th test Tv is dropped down the $T(v)(\alpha)$; $v = 1, 2, 3, ..., V$. Let $n_{ij}^{(v)}$ be the number of j-th class observations in T_v that are classified as being from the i-th class, $i, j = 1, 2, 3, ..., K, \quad v = 1, 2, 3, ..., V$.

Because,

$$D = \cup T_v \tag{3.13}$$

is a disjoint sum, the total number of j-th class observations that are classified as being from the i-th class is;

$$\sum_{i,j}^{v} \alpha \qquad \text{where } i, j = 1, 2, 3, ..., K$$

If n_j is the number of observations in D that belong to the j-th class, $j = 1, 2, 3, ..., J$, and assuming equal misclassification for all classes, then for a given value of α,

$$R^{CV/V} = T(\alpha) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}\alpha}{n}$$

Is the misclassification rate over D, where $T(\alpha)$ is a minimizing subtree of T_{max}. The final step in this process is to find the right sized subtree. For different α values, $R^{CV/V}$ is evaluated. If for a sequence of values k, corresponding cross-validation error of the minimizing subtree $T(\alpha) = T_k$ is given by;

$$R^{CV/V}(T - k) = RCV = V = T(\alpha k)$$

Then, the best-pruned subtree $T\alpha$ is selected by the rule,

$R^{CV/V}(T*) = \min\ kR^CV = \text{V}= (T_{\alpha_k})$ and $R^{CV} = V = (T_{\alpha_k})$ is used as its estimated misclassification rate.

Chapter 4

Data Analysis and Results

Introduction

This chapter basically deals with the exploratory, inferential analysis and modeling of the data collected. Claims paid data from the period 2012 to 2014 was used to develop a panel data of policyholders, whether the policyholder has made a claim or not. The total number of policyholders that were considered is One Thousand Five Hundred and Thirty-Nine (1,539). Risk characteristics of policyholders was recorded against their claim status, yes or no. The categorical variables like make of vehicle, Gender, usage used in the analysis are fixed and do not vary overtime. However, the numerical variables like age of policyholder and age of vehicle changed overtime.

Exploratory Analysis of Study Variables

The target variable that was used is the question “has the policyholder made a claim?" The response was categorized as Yes or No. Below is the summary of the data which is limited only to the training dataset.

Distribution of Variables

The independent variables considered in the analysis were numerical and categorical variable. Categorical variables have two major scales; nominal and ordinal scale. In this thesis all the variables that were used were nominal.

From Table 4.1, Policyholders who answered that they have made at least

one claim from represents 59% and those who answered that they have not made claim represents 41%. Also, from Tab 4.2, about 32% of the policyholders were in the corporate group and about 68% belong to the Individual group.

Table 4.1: Summary - Claim Status

CLAIM?	FREQ.	PROP.
YES	662	43%
NO	866	57%
TOTAL	1528	100%

Table 4.2: Summary - Policyholder Group

GROUP	FREQ.	PROP.
CORP	487	32%
IND.	1041	68%
TOTAL	1528	100%

Table 4.3: Summary - Vehicle Usage?

USAGE	FREQ.	PROP.
COMM.	84	5%
PRIV.	1444	95%
TOTAL	1528	100%

From Table 4.4, about 25% of Policyholders are less than 31 years, 50% of Policyholders are less than 38 years, and 75% less than 43 years. Also, the Median and the Mean Policyholder age are 38 years each. This means that the age of policyholders is symmetrical about the mean.

Moreover, Table 4.4 shows that on the whole, 25% of the Policyholders have their vehicles aged below 7 years and 50% have their vehicles aged below 9 years. However, as compared to the Age of Policyholders, the mean vehicle age and the median age of a vehicle are 12.25 years and 9 years respectively. This shows that the age of vehicles is asymmetrical about the mean.

Table 4.4: Summary Statistics

	Policyholder Age	Vehicle Age
Minimum	19	0
1st Quartile	31	7
Median	38	9
Mean	38	12.25
3rd Quartile	43	15
Maximum	64	34

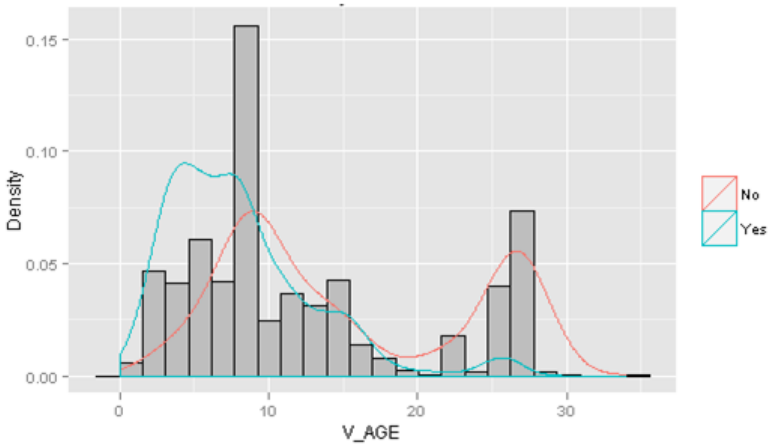

Figure 4.1: Superposition of the kernel density on the age of Vehicle

Figure 4.1 and Figure 4.2 shows how the kennel density for vehicle age and policyholder age is superimposed on the histogram for policyholders claim status. In studying the age of policyholders and Age of Vehicle which have been identified as major risk factor, it can be observed that policyholders who have made at least one claim have their age being bi-modal whilst those who have not made a claim have their age being highly skewed.

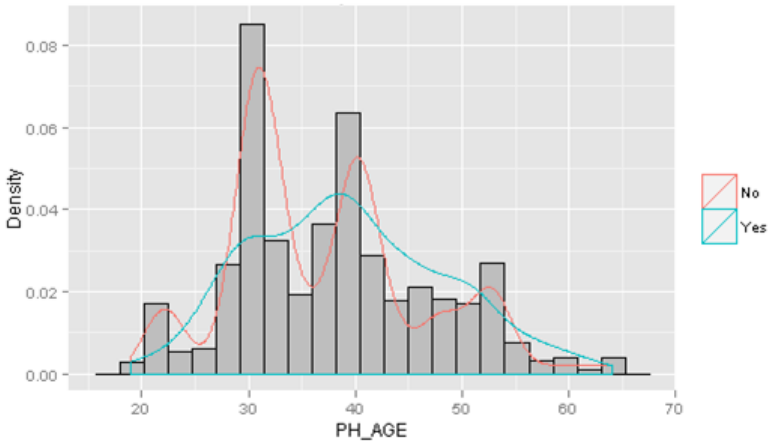

Figure 4.2: Superposition of the kernel density on the age of the Policy Holder

The Classification Tree

A) Analysis On The Individual Customers

From Figure 4.3, the root node with 728 observations was split into whether V_AGE >= 21 (229 with 21 yes and 208 No) or V_AGE <21(499, with 174 No and 325 Yes).

Further, the node with 229 observations was split into PH_AGE >= 30.5 with total observation 216 (11 yes, 205 No) and PH_AGE<=30.5 with 13 observations (11 No and 3 Yes). The data consist of 728 observations (346 experienced claim within the period and 382 did not experience claim within the period. i.e if the root node was used as a model it will always declare that there would be no claim with 52.47% probability.

Node 4 predicts, for the moment that there would be no claim for an individual claimant who is aged above 301/2 years and vehicle age is greater than or equal to 21 would have no claim with 94%.

Whilst Node 5 also predicts that an individual policyholder who is aged less than 30 years who owns a vehicle aged less than 21 would result in a claim with 77% chance. Thus Node 4 is classified as terminal nodes. According to

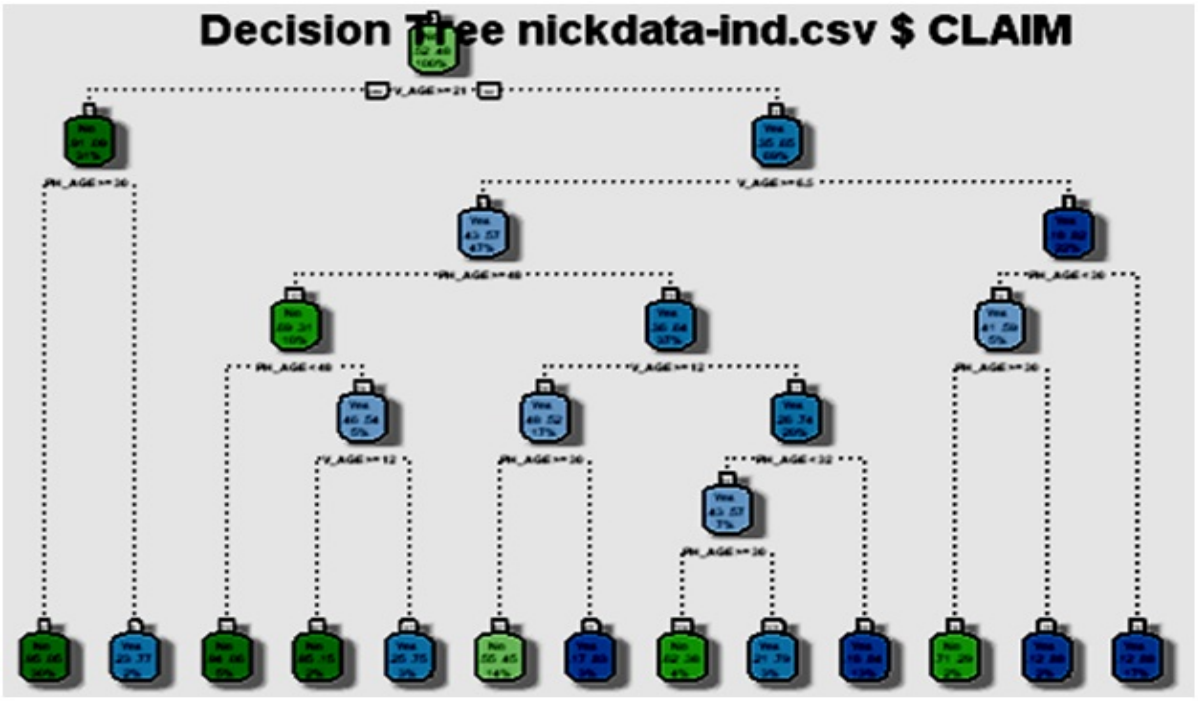

Figure 4.3: Decision Tree For Individual

the same Figure 4.3, the node with 499 observations categorized private insured customers into V_age>=6.5 (341 with 145 No, 196 yes) and V_age<6.5 (158 with 29 No, 129). The node with 341 observations was split by PH_Age>=48 (70 with 22 yes, 48 No). and PH_age<48 (271 with 97 No, 175 Yes). In furtherance to the classification the node with 70 observations was split by PH_age<=48.5 (33 with 31 No, 2 yes) and PH_age>48.5 (37, with 17 No, 20 Yes). Node 24 predicted with 95% probability that an insured individual customer aged over 48 years whose vehicle is above 6.5 years would not get involve with a claim. Thus node 24 is a terminal node.

The node with 37 data observed split into V_age>=12 (13 with 2 Yes, 11 No) and V_age<12 (24 with 6 No, and 18 Yes).

Therefore, **Node 50**, being a terminal node predicted that an individual customer who is at least 48.5 and the vehicle aged at least 12 years has 85% chance of not making a claim, whilst **Node 51** predicts that 75% chance of not making a claim for an individual who is less than 48.5 years driving a vehicle of less than 12 years.

The node with 271 observations split into v_age< 12 (124, with 59 No, 65 Yes) and v_age<=12 (147 with 59 No, 88 Yes).

The node with 124 data observations further split into Ph_age>=30 (100, 45 yes, 55 No) and Ph_Age< 30 (24 with 4 No, 20 Yes).

The CP column lists the complexity parameter (α) at each stage of the tree growing process; and the xerror shows the rate of cross validation error; the xstd measures the standard deviation of the xerror standard. The relative error, rel error keeps le decreasing as the tree grows bigger and starts rising to some point (see Table 4.5)

Table 4.5 provides a brief summary of the overall fit of the model. The table is

Table 4.5: Cross Parameter table for Individual

No.	CP	nsplit	rel error	xerror	xstd
1	0.436416	0	1	1	0.038943
2	0.037572	1	0.56358	0.56647	0.034589
3	0.020231	3	0.48844	0.49133	0.032991
4	0.017341	4	0.46821	0.49711	0.033125
5	0.014451	6	0.43353	0.50289	0.033257
6	0.010116	8	0.40462	0.49133	0.032991

printed from the smallest tree (nsplits= 0) to the largest tree (nsplit=12). The number of nodes of a Tree is given by (1+nsplit).

From table 4.3, the (1-SE) rule yields a $\min(CV^{err} + SE) = 0.49133 + 0.033125 = 0.5246$.

Table 4.6: Confusion Matrix-Individual

Actual	Predicted		
	No	Yes	Error
No	0.48	0.07	0.13
Yes	0.11	0.34	0.24
Overall Error:			0.1783
Class Error:			0.177

From Table 4.6, it is clear that there is 48% chance that there would be No Claim by a private individual also known as 'the true negative' and 34% chance that there would be Claim by an individual in the coming year. An average class error of 0.170067, shows that the model is a good model at least based on the data available.

B) Analysis on Corporate clients

The training dataset consists of 340 observations (123 Yes, 217 No). This implies that corporate policyholders in general have 36% chance of making a claim and a 64% chance of not making a claim. The classification tree is displayed in Figure 4.4, where the entropy measure was used as the impurity function for splitting. There is 1 split and 2 terminal nodes.

From the classification tree, it can be predicted with 13% probability that a corporate insured customer with vehicle aged more than 8.5 years will make a claim in the coming year. However, a corporate insured customer vehicle that is aged less than 8.5 years has 81% probability of making a claim in the next year.

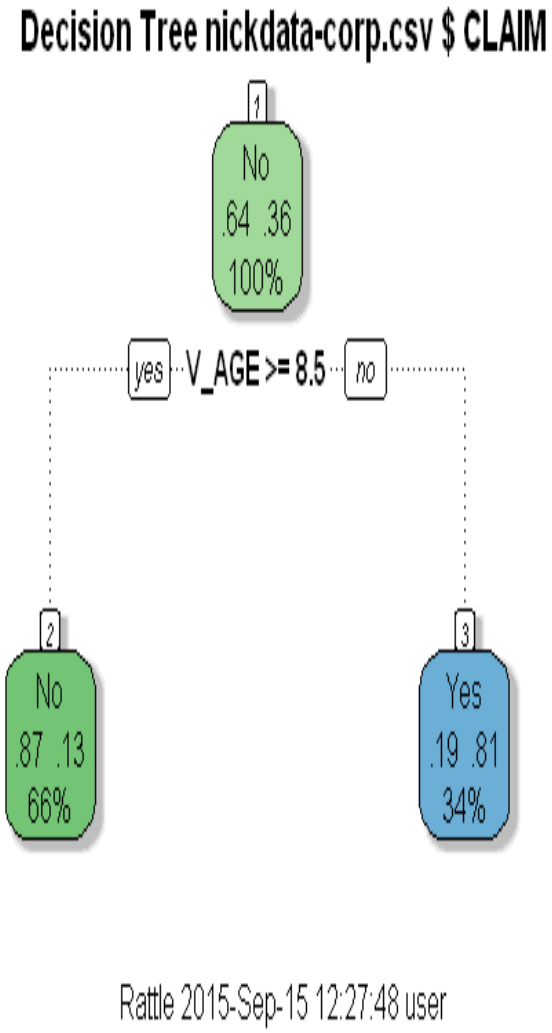

Figure 4.4: Classification Tree for Corporate

B1) Misclassification rate for Corporate

In Table 4.7 out of the 340 observations, the classification algorithm misclassified 34 of persons who had not made claim as having made claim and 41 of those who has made claim and not going to make claim.

So From the formulae for the resubstitution error,

$$R^{re}(T) = \frac{34 + 41}{340} = \frac{75}{340} = 0.22$$

Table 4.7: Confusion Matrix-Corporate

Actual	Predicted		
	No	Yes	Error
No	0.58	0.07	0.11
Yes	0.08	0.26	0.24
Overall Error:			0.154
Class Error:			0.168

Table 4.8: Cross Parameter table for Corporate

No.	CP	nsplit	rel error	xerror	xstd
1	0.58537	0	1	1	0.07203
2	0.01	1	0.04146	0.041463	0.05329

From Table 4.7 and Tab 4.8 it shows that corporate policyholders that did not make any claim in the current year would not make a claim in the coming year with 58% probability. Also there is a 7% chance that a corporate customer who did not make a claim in the current year would make a claim in the coming year. With an overall error of approximately 15%, it shows the model is a good prediction at least given the data at our disposal data.

C) Analysis on the Overall Data

The data consists of 1069 claimants (461 No, 608 Yes).

From Figure 4.5, the Root Node with 1069 observations was split by V_AGE>=8.5 (670 with 165 Yes, 445 No) and V_AGE<8.5 (399, with 103 No, 296 Yes). The node with 670 observations was split by V_AGE>=26(188 with 8 Yes, 180 No) an V_AGE<26 (482 with 157 Yes, 325 No).

Node 4 with 95% accuracy predicts that a person whose vehicle age >= 26 has sure probability of making No claim in the coming. Now for the node with 482 observations the algorithm split the data by Group=Corp (222 with 26 yes, 118 No) and Group =Individual (260 with 129 No, 131 Yes). **Node 10** is a terminal node and predicts with 88.28% probability that a corporate claimant whose vehicle age is less than 26 years would not likely report a claim in the

coming year.

According to Figure 4.5For the node with 144 observation it is split by PH_AGE>=47.5 (48 with 19 Yes, 29 No) and PH_AGE<47.5 (96 with 24 No, 72 Yes).

The node with 260 observations was split by PH_AGE>=27.5(233 with 106 Yes, 127 No) and PH_AGE < 27.5 (27 with 2 No, 25 Yes). Node 23 is a terminal node and predicts with 92.6% probability that a policyholder who is a private individual and aged less than 27.5 would make a claim next year. The node with 233 observations is split by PH_AGE>=21 (32 with 3 yes, 29 No) and PH_AGE<21 (201 with 98 No, 103 Yes). Therefore **Node 44** is chosen as a terminal node and predicts with 98% probability that a corporate policyholder aged above 21 years would not make a claim in the coming.

Table 4.9: Cross Parameter Table –Overall Data

	CP	nsplit	rel error	xerror	xstd
1	0.418655	0	1	1	0.035125
2	0.016631	1	0.58134	0.58134	0.030739
3	0.015907	7	0.51193	0.51193	0.029416
4	0.010846	10	0.47289	0.47289	0.029576
5	0.01	11	0.48156	0.48156	0.028769

Table 4.10: Confusion Matrix-All Data

Actual	Predicted		
	No	Yes	Error
No	0.543	0.11	0.21
Yes	0.11	0.34	0.25
Overall Error:			0.227
Class Error:			0.228

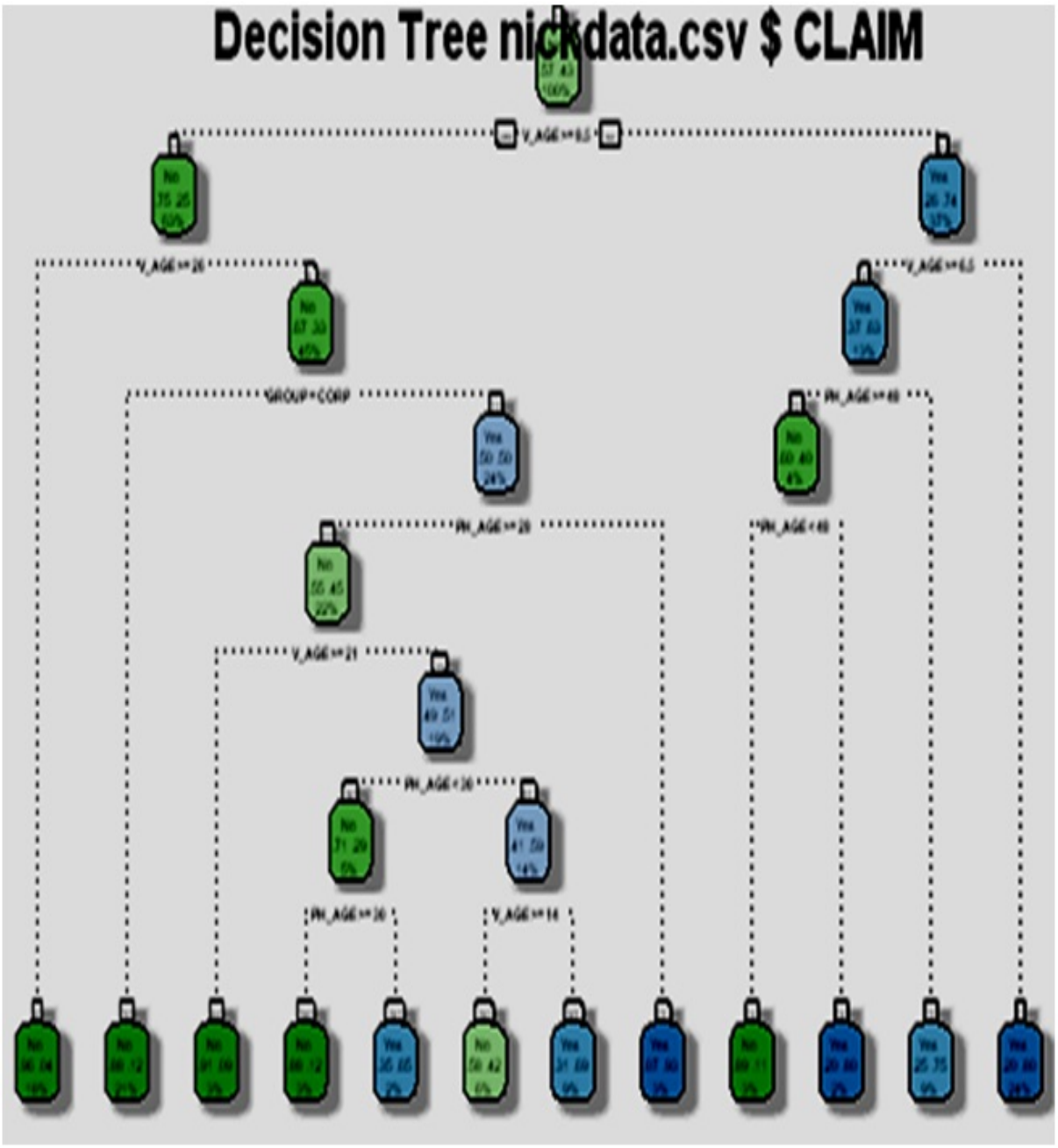

Figure 4.5: Classification Tree For All Data

Chapter 5

Discussion, Conclusions and Recommendations

Discussion

From Table 4.1 it can be observed that generally, the number of policyholders that made at least one claim represents 43% of the entire policyholders of the company and those that did not make a claim represents 57%. This supports and backs the general view of the public that most persons do not make claim to the insurance company. However, that can not be concluded to say that the effect of claims do not affect the companies growth. Since the principle of insurance operates by the law of large numbers, it may be that even though a lot of policyholders comes to the "pool", just a few of them get register claimable event(s). Therefore an assessment of individual claim size could help to ascertain the proportions above. For instance third party injury claim liability is unlimited, which means though the count may be 1 its impact could be felt when it comes to modeling with the claim size .

A critical examination of Table 4.6, shows that the model predicts that if a policyholder owns his/her own car and did not make a claim in 2014, then there is 48% chance that he will not make a claim in the coming year and 7% otherwise. However, for a person who makes a claim in the current year i.e. 2014, there is a 34% probability that he will make a claim in 2015.

Analysis of corporate policyholders shows a similar trend as depicts from that of individual as shown in Table 4.7. Thus, there is a 58% chance that a corporate policyholder, who did not make a claim in the current year, will also

not make a claim in the next year as against 7% for those who did not make a claim this year but will make a claim next year.

Similarly, from Table 4.7, for corporate policyholders who made a claim the current year, there is a 8% chance that there will be no claim as against 26% that they will make a claim in the next year.The overall prediction of claim shows that the propensity for a policyholder to claim depends very much on the loss history of that claimant.

Also from Table 4.3, commercial vehicles impact on claim reporting was quite negligible i.e. 5% as against 95% for vehicles that are used for privately used. This could be attributed to the fact that most commercial drivers feel reluctant to report claims, lacks adequate knowledge on insurance, inability to compile claim documentations, whilst private persons who owns their vehicles have adequate knowledge of claim processes, have tendency to challenge liability of insurance policies, etc. However the claim size of a commercial vehicle could have a significant impact on the insurance company's funds than a lot more private car.

In addition, the foregone data analysis made it clear that age of policyholders and age of vehicles are the most factors that affect the propensity of a policyholder to make a claim, and it varies among the two groups that were identified i.e. Individual and Corporate policyholders with the following key observations(From Table 4.4 and Figure 4.3):

a) Whilst the age of the policyholder is symmetrical about the mean, the age of the vehicle is asymmetrical.

b) Policyholders aged between 30 years and 40 years have a high propensity to make claim, and very peaked at age 30.

c) Individual policyholders aged more than 48 years has a low rate of reporting a claim. This could be attributed to various reasons; low education level, inadequate information on insurance claim, "tiring" claim processes, etc.

Privately used vehicles have more potential to make claims than commercial vehicles. This is as a result of the time consuming claim process, the "knock-for-knock" agreement that commercial drivers do at time of accident and the fear of police for lack of vehicle documents among commercial vehicle users. Also most private vehicle owners in Ghana have high level of education and insurance awareness, hence are able to read insurance contracts to understand, follow up claim processes to its logical conclusion and provide legal arguments where needed.

From Figure 4.3 and Table 4.4 young adults between the ages of 19 years to 36 years have high propensity to make claim than older folks. This can be attributed to the fact that younger adults have more potential and expose themselves to the risk on the road, due to inexperience, non observance of road safety measures etc.

Vehicles aged between 0 to 8 years make lots of claim as compared to vehicles above 8 years. This is obviously the fact that owners of new and less older vehicles belongs to the elite class who have more knowledge in insurance and ready to follow the claim processes to its logical conclusion as compared to those of older vehicles.

Vehicles that are own by individuals have high claim reporting rate than those that belongs to a corporate bodies.

Conclusions

Among the predictor variables that were used in the study, to predict claim, the age of the vehicle and the age of the policyholder was selected as a major predictor variable as compared to the other variables.

For the purpose of predictive modeling, there are two major classes that policyholders can be put into; corporate and individual policyholders.

Individual policyholders make marginally more claim as compared to corporate customers. However, the policyholder age and vehicle age has more effect on the individual class than on the corporate class.

Generally, there is a greater chance for individuals or corporate bodies who make a claim in the current year to make a claim in the ensuing year.

The low error margin of the prediction shows that the model is well validated and suitable for the prediction of future claims given the current data and risk characteristics.

Recommendations

It is recommended that the study includes lots of other potential risk factors, such as NCD level, education level, etc. so as to be able to improve the the predictive powers of the model with greater certainty.

The insurance industry must undertake public education on insurance products to increase awareness of claim processes and procedure among less educated and the low income policyholders.

Appropriate insurance premium rate should be applied to private vehicle owners that are within the age of 19 years to 36years.

Moreover, I highly recommend that premium reduction strategy be adopted as part of the rating scheme for the class of policyholders that depicts less claim record.

Finally, I recommend that further and more in-depth classification analysis be carried out with funding from stakeholders which would include data from other insurance companies to improve the predictive power of the model and assess its application as an alternative or addendum to premium pricing mechanism.

Reference

Austin, P. C. (2008). R and S-PLUS produced different classification trees for predicting patient mortality. Journal of clinical epidemiology, 61(12), 1222-1226.

B. Efron and R. Tibshirani. Cross-validation and the Bootstrap: Estimating the Error Rate of a prediction rule. Technical report, Standford University(1995).

Banker, R. D., & Hansen, S. C. (2002). The adequacy of full-cost-based pricing heuristics. Journal of Management Accounting Research, 14(1), 33-58.

Berry, M. J., & Linoff, G. (1997). Data mining techniques: for marketing, sales, and customer support. John Wiley & Sons, Inc..

Breiman L. (1996). Bagging Predictors, Machine Learning, 24, 123-140.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Buckinx, W., Moons, E., Van den Poel, D., & Wets, G. (2004). Customer-adapted coupon targeting using feature selection Expert Systems with Applications, 26(4), 509-518.

Choi, D. H., Ahn, B. S., & Kim, S. H. (2005). Prioritization of association rules in data mining: Multiple criteria decision approach. Expert Systems with Applications, 29(4), 867-878.

Dal Pozzolo, A., Moro, G., Bontempi, G., & Le Borgne, D. Y. A. (2010). Comparison of Data Mining Techniques for Insurance Claim Prediction (Doctoral dissertation, PhD thesis, University of Bologna).

Duodu, F.K & Amankwaa T. (2011). An analysis and assessment of customer satisfaction with service quality in insurance industry in Ghana

Fadun, O. S. (2013). Insurance, A Risk Transfer Mechanism: An Examination Of The Nigerian Banking Industry.

Giraud-Carrier, C., & Povel, O. (2003). Characterising data mining software. Intelligent Data Analysis, 7(3), 181-192.

http://www.312analytics.com/decision-trees-vs-neural-networks/(14/03/2016)

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning:A conditional inference framework, Journal of Computational and Graphical Statistics 15(3), 651; 674. 241, 242, 266.

Hölmstrom, B. (1979). Moral hazard and observability. The Bell journal of economics, 74-91.

Izenman, A. J. (2008). Modern Multivariate Statistical Techniques. *Springer Science Business Media*, LLC, 233 New York, NY 10013, USA.

J., Harnos, A., Schrott, A., and Kabai, P. (2010).Tree-based methods as an alternative to logistic regression in revealing risk factors of cribbiting in horses.Journal of Equine Veterinary Science Vol 30, No 1 (2010), 30:2126.

Kajungu, D. K., Selemani, M., Masanja, I., Baraka, A., Njozi, M., Khatib, R., ... & Speybroeck, N. (2012). Using classification tree modelling to investigate drug prescription practices at health facilities in rural Tanzania. Malaria journal, 11(1), 1.

Kaiser, G. E. (1988). Database Support for Knowledge-Based Engineering Environments. IEEE Expert , 18-32.

Kahneman, D., & Slovic, P. (1982). Amos Tversky, eds. 1982.

Kass G.V (1980). An Exploratory Technique for investigating Large Quantities of Categorical Data. Applied Statistics, 29: -119-127,1980. Nagy, K., Reiczigel,

Loh, W. Y., & Shih, Y. S. (1997). Split selection methods for classification trees. Statistica sinica, 815-840.

Mayhew, D. R., Simpson, H. M., & Pak, A. (2003). Changes in collision rates among novice drivers during the first months of driving. Accident Analysis & Prevention, 35(5), 683-691.

May, M., Tranter, P. J., & Warn, J. R. (2011). Progressing road safety through deep change and transformational leadership. Journal of Transport Geography, 19(6), 1423-1430.

Mehr, R., & Hedges, B. (1963). Risk Management in the Business Enterprise (Homewood, IL: Irwin)

Meldrum, D. (2000). Country risk and foreign direct investment. Business economics, 35(1), 33-40.

Moore, A. W. (2001). Decision Trees. Professor School of Computer Science Carnegie Mellon University http://www. autonlab. org/tutorials/dtree18. pdf.

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.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: a survey. Neural Networks, IEEE Transactions on, 13(1), 3-14.

Murphy, K. P., M. J. Brockman, and P. K W Lee (2000). Using generalized linear models to build dynamic pricing systems for personal

lines insurance.Casualty Actuarial Society Winter Forum Winter 2000,107-140.

Nagy, K., Reiczigel, J., Harnos, A., Schrott, A., and Kabai, P. (2010). Tree-based methods as an alternative to logistic regression in revealing risk factors of cribbiting in horses. *Journal of Equine Veterinary Science* Vol 30, No. 1, 30:21 – 26.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.

Ritschard, G. (2006). Computing and using the deviance with classification trees. In COMPSTAT 2006-Proceedings in Computational Statistics (pp. 55-66). Physica-Verlag HD.

Roy, A. and Roy, P.G., (1994), Despite Past Debacles, Predicting Sovereign Risk Still Presents Problems, Commercial Lending Review, Summer, 9 (3), 92-95

R. Sipulskyte (2012). Development of a Vehicle Classification Scheme for New Zealand Insurance Co.

Shavell, S. (1979). On moral hazard and insurance (pp. 280-301). Springer Netherlands.

Therneau T., Atkinson E.(1997) An Introduction to Recursive Partitioning Using the RPART Routines, Mayo Foundation, 1997.

Tryfos, P. (1980). On classification in automobile insurance. The Journal of Risk and Insurance, 47(2), 331-337.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining (Vol. 1). Boston: Pearson Addison Wesley.

Therneau, T. M., & Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines.

Volume Information. (1980). Volume Information. The Journal of Risk and Insurance, 47(4), 780–783. Retrieved from http://www.jstor.org/stable/252282 (accessed on 16/03/2016)

Wenzel, T. P., & Ross, M. (2005). The effects of vehicle model and driver behavior on risk. Accident Analysis & Prevention, 37(3), 479-494.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

Xhemali, D., Hinde, C. J., & Stone, R. G. (2009). Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages.

Yeo, A., K. Smith, J. Robert, R. Willis and M. Brooks.(2001) "Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry." International Journal of Intelligent Systems in Accounting, Finance & Management 10, no. 1:39-50. Volume Information. (1980).

Appendix

APPENDIX A

Table 5.1: Type of Variables

Variable	Type	No. of Classes
Policyholder Age	Numeric	None
Vehicle Age	Numeric	None
Make of Vehicle	Ordinal	58
Gender	Categorical	3
Vehicle Use	Categorical	2

APPENDIX B

Summary of the Decision Tree model for Classification (built using 'rpart'): Corporate

```
n= 340

Legend:

node), split, n, loss, yval, (yprob) * denotes terminal node

1) root 340 123 No (0.6382353 0.3617647)

2) V_AGE>=8.5 224  29 No (0.8705357 0.1294643) *

3) V_AGE< 8.5 116  22 Yes (0.1896552 0.8103448) *


Classification tree:

rpart(formula = CLAIM ~ ., data = crs$dataset[crs$train, c(crs$input,

crs$target)], method = "class", parms = list(split = "information"),

control = rpart.control(usesurrogate = 0, maxsurrogate = 0))


Variables actually used in tree construction:
```

```
[1] V_AGE

Root node error: 123/340 = 0.36176
n= 340

CP   nsplit   rel   error     xerror    xstd
1    0.58537   0   1.00000   1.00000   0.072034
2    0.01000   1   0.41463   0.41463   0.053529
```

APPENDIX C

Summary of the Decision Tree model for Classification (built using 'rpart'): Private

```
Legend
node:), split, n, loss, yval, (yprob) * denotes terminal node

1) root 728 346 No (0.52472527 0.47527473)
2) V_AGE>=21 229  21 No (0.90829694 0.09170306)
4) PH_AGE>=30.5 216  11 No (0.94907407 0.05092593) *
5) PH_AGE< 30.5 13   3 Yes (0.23076923 0.76923077) *
3) V_AGE< 21 499 174 Yes (0.34869739 0.65130261)
6) V_AGE>=6.5 341 145 Yes (0.42521994 0.57478006)
12) PH_AGE>=47.5 70  22 No (0.68571429 0.31428571)
24) PH_AGE< 48.5 33   2 No (0.93939394 0.06060606) *
25) PH_AGE>=48.5 37  17 Yes (0.45945946 0.54054054)
50) V_AGE>=11.5 13   2 No (0.84615385 0.15384615) *
51) V_AGE< 11.5 24   6 Yes (0.25000000 0.75000000) *
13) PH_AGE< 47.5 271  97 Yes (0.35793358 0.64206642)
```

```
26) V_AGE>=11.5 124  59 Yes (0.47580645 0.52419355)
52) PH_AGE>=29.5 100  45 No (0.55000000 0.45000000) *
53) PH_AGE< 29.5 24   4 Yes (0.16666667 0.83333333) *
27) V_AGE< 11.5 147  38 Yes (0.25850340 0.74149660)
54) PH_AGE< 31.5 53  23 Yes (0.43396226 0.56603774)
108) PH_AGE>=29.5 29  11 No (0.62068966 0.37931034) *
109) PH_AGE< 29.5 24   5 Yes (0.20833333 0.79166667) *
55) PH_AGE>=31.5 94  15 Yes (0.15957447 0.84042553) *
7) V_AGE< 6.5 158  29 Yes (0.18354430 0.81645570)
14) PH_AGE< 30.5 34  14 Yes (0.41176471 0.58823529)
28) PH_AGE>=29.5 17   5 No (0.70588235 0.29411765) *
29) PH_AGE< 29.5 17   2 Yes (0.11764706 0.88235294) *
15) PH_AGE>=30.5 124  15 Yes (0.12096774 0.87903226) *

Classification tree:
rpart(formula = CLAIM ~ ., data = crs$dataset[crs$train, c(crs$input,
crs$target)], method = "class", parms = list(split = "information"),
control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] PH_AGE , V_AGE

Root node error: 346/728 = 0.47527
```

APPENDIX D

Summary of the Decision Tree model for Classification (built using 'rpart'): ALL DATA

```
n= 1069

node), split, n, loss, yval, (yprob) * denotes terminal node


1) root 1069 461 No (0.56875585 0.43124415)
2) V_AGE>=8.5 670 165 No (0.75373134 0.24626866)
4) V_AGE>=25.5 188   8 No (0.95744681 0.04255319) *
5) V_AGE< 25.5 482 157 No (0.67427386 0.32572614)
10) GROUP=CORP 222  26 No (0.88288288 0.11711712) *
11) GROUP=IND. 260 129 Yes (0.49615385 0.50384615)
22) PH_AGE>=27.5 233 106 No (0.54506438 0.45493562)
44) V_AGE>=21 32   3 No (0.90625000 0.09375000) *
45) V_AGE< 21 201  98 Yes (0.48756219 0.51243781)
90) PH_AGE< 30.5 51  15 No (0.70588235 0.29411765)
180) PH_AGE>=29.5 34   4 No (0.88235294 0.11764706) *
181) PH_AGE< 29.5 17   6 Yes (0.35294118 0.64705882) *
91) PH_AGE>=30.5 150  62 Yes (0.41333333 0.58666667)
182) V_AGE>=13.5 59  25 No (0.57627119 0.42372881) *
183) V_AGE< 13.5 91  28 Yes (0.30769231 0.69230769) *
23) PH_AGE< 27.5 27   2 Yes (0.07407407 0.92592593) *
3) V_AGE< 8.5 399 103 Yes (0.25814536 0.74185464)
6) V_AGE>=6.5 144  53 Yes (0.36805556 0.63194444)
12) PH_AGE>=47.5 48  19 No (0.60416667 0.39583333)
24) PH_AGE< 48.5 28   3 No (0.89285714 0.10714286) *
25) PH_AGE>=48.5 20   4 Yes (0.20000000 0.80000000) *
13) PH_AGE< 47.5 96  24 Yes (0.25000000 0.75000000) *
```

```
7) V_AGE< 6.5 255  50 Yes (0.19607843 0.80392157) *

Classification tree:
rpart(formula = CLAIM ~ ., data = crs$dataset[crs$train, c(crs$input,
crs$target)], method = "class", parms = list(split = "information"),
control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] GROUP,  PH_AGE, V_AGE
```