



Application of predictive data mining technique to predict Claim
Cost of Risk Items under Motor Class of Business: The Case of
Awash Insurance Company S.C.

A Thesis Presented

by

Deresse Berhanu Tola

to

The Faculty of Information

of

St. Mary's University

In Partial Fulfillment of the Requirements

For the Degree of Master of Science

in

Computer Science

February 22, 2020

ACCEPTANCE

Application of predictive data mining technique to predict Claim Cost of Risk

Items under Motor Class of Business: The Case of Awash

Insurance Company S.C.

By

Deresse Berhanu Tola

Accepted by the Faculty of Informatics, St. Mary's University, in partial

fulfillment of the requirements for the degree of Master of Science in

Computer Science

Thesis Examination Committee:

Dr. Asrat Mulatu

Internal Examiner

Signature

Date

Dr. Temtim Assefa

External Examiner

Signature

Date

Dr. Getahun Semeon

Dean, Faculty of Informatics

Signature

Date

February 7, 2020

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Deresse Berhanu Tola

Full Name of Student

Signature

Addis Ababa

Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Dr. Getahun Semeon

Full name of Advisor

Signature

Addis Ababa

Ethiopia

February 7, 2020

Acknowledgments

Words can never be enough to thank the almighty God who helped me to finish this thesis

I would like to thank my Advisor, Dr. Getahun Semeon for his guidance, encouragement and kind personality.

I would like to thank friends, branch underwriters and branch managers, claim offices and claim managers of Awash insurance company for their support and encouragement, which gave me strength to successfully complete this work

Finally, I would like to thank my darling wife Mrs, Aster Tesfaye for valuable support during my study.

Table of Contents

Acknowledgments.....	i
List of Acronyms	vi
List of Figures.....	vii
List of Tables	ix
Abstract.....	x
CHAPTER ONE:	1
1 INTRODUCTION.....	1
1.1 Background of the Study.....	1
1.2 Motivation	3
1.3 Statement of the problem	3
1.4 Research Question.....	9
1.5 Objective of the Study.....	9
1.5.1 General objective	9
1.5.2 Specific objective.....	9
1.6 Scope of the Study.....	9
1.7 Limitations	10
1.8 Significance of the study	10
CHAPTER TWO:	11
2 REVIEW OF LITERATURES.....	11
2.1 Introduction	11
2.1.1 Introduction to Insurance	11
2.2 Development of Insurance in brief.....	12
2.3 Principles of Insurance	12
2.4 Ethiopian Insurance History in Brief	13

2.5	Motor Insurance in Ethiopia.....	15
2.6	Risk.....	16
2.6.1	Introduction to Risk	16
2.6.2	Risk and Insurance	17
2.7	Data Mining.....	19
2.7.1	Introduction to Data Mining	19
2.7.2	Acquiring New Customers.....	22
2.7.3	Retaining Existing Customers.....	22
2.7.4	Classification: Databases Segmentation	22
2.7.5	Data Mining Techniques.....	23
2.8	Predictive models	25
	CHAPTER THREE:	27
3	RESEARCH METHODOLOGY	27
3.1	Introduction	27
3.2	The Data Mining Process	27
3.2.1	Understanding of the problem	27
3.2.2	Understanding of the data	28
3.2.3	Preparation of the data:	28
3.2.4	Data modeling:.....	28
3.2.5	Evaluation of the discovered knowledge:	28
3.2.6	Use of the discovered knowledge.	29
3.3	Data Mining Process Modeling.....	29
3.3.1	Knowledge Discovery Databases (KDD).....	29
3.3.2	Sample, Explore, Modify, Model, and Access (SEMMA).....	30
3.3.3	Cross-Industry Standard Process for (CRISP-DM) processing model.....	31

3.3.4	Comparison of KDD, SEMMA and CRISP-DM.....	33
3.4	Evaluation.....	37
3.5	Research design.....	39
3.5.1	Data collection	39
3.5.2	Data Pre-Processing	41
3.5.3	Data integration.....	42
3.5.4	Data formatting	42
3.5.5	Maintaining Balances.....	43
3.5.6	Attribute Selection	43
3.6	Data Transformation	45
3.7	Framework for Guiding Data Mining Tasks	50
3.7.1	Business understanding.....	50
3.7.2	Data Understanding	51
3.7.3	Data Preparation.....	52
3.7.4	Modeling	52
3.8	Data Visualization	53
	CHAPTER FOUR:	63
4	EXPRIMENT AND DISCUSSION OF THE OUTPUT	63
4.1	Introduction	63
4.2	Experiment	63
4.2.1	Support Vector Machine (SVM) modeling.....	63
4.3	Dividing the dataset into training and test set	66
4.4	Naïve Bayes Model	70
4.4.1	Data set splitting into training data and testing data.....	71
4.4.2	Naïve Bayes model building.....	72

4.5	Logistic Regression Modeling	75
4.5.1	Logistic Regression model building	77
4.6	Summary of the findings	82
4.7	Deployment	84
4.7.1	Prototype	84
CHAPTER FIVE:		88
5	CONCLUSION AND RECOMMENDATION	88
5.1	Introduction	88
5.2	Conclusion.....	88
5.3	Recommendations and future work.....	89
Annex 1		97
Annex 2.....		99
Annex 3.....		101
Annex 4.....		102

List of Acronyms

PN	Policy Name
PC	Policy Code
PA	Premium Amount
CA	Claim Amount
CH	Claim History
RN	Risk Name
VM	Vehicle Make
PoV	Purpose of the Vehicle
AoV	Age of the Vehicle
CC	Carrying Capacity of the vehicle
PRM	Premium Calculation Rate
CSV	Comma Separated Values
DM	Data Mining
CRISP	Cross Industry Standard Process
BPNN	Back Propagation Neural Network
KDD	Knowledge Discovery Database
SEMMA	Sample, Explore, Modify, model, Access

List of Figures

Figure 2. 1 : KM Technologies Integrated KM Cycle (Source from Dalkir, K., 2005).....	20
Figure 3. 1 : The Five Stages of KDD, Source: Xiao Zhu, 2017	30
Figure 3. 2: SEMMA Cycle.Source: Xiao Zhu, 2017.....	31
Figure 3. 3 : CRISP-DM Process Model, Source: Shafique, &Qaiser, 2017.	33
Figure 3. 4 : Comparison of Data Mining process modes	35
Figure 3. 5: KDnuggets poll results (Piatesky, 2014), Source: Xiao Zhu, 2017	36
Figure 3. 6: Policy type distribution	54
Figure 3. 7 : Premium collected distribution.	55
Figure 3. 8: Claim paid distribution.....	56
Figure 3. 9: Claim history distribution.....	57
Figure 3. 10 : Risk Items distribution with number representation	57
Figure 3. 11 : Risk Items distribution with actual risk items name	58
Figure 3. 14 : Vehicle- make graphical representation	58
Figure 3. 15 : Number of vehicles by purpose.....	60
Figure 3. 16 : Number of vehicles by age of the vehicles under the experiment.....	61
Figure 3. 17 : Number of Vehicles by Carrying Capacity of the vehicles under experiment.....	62
Figure 4. 1 : Dataset for SVM model snap shot.....	64
Figure 4. 2 : Dataset structures for SVM model snap shot	65
Figure 4. 3 : Dataset summary for SVM model snap shot.....	66
Figure 4. 4: Partitioning the dataset snap shot	66
Figure 4. 5 : Dataset Partitioning Resultfor SVM model snap shot.....	67
Figure 4. 6 : Building SVM model snap shot	67
Figure 4. 7 : Prediction and model accuracy for SVM model snap shot	68
Figure 4. 8 : Confusion Matrix result for SVM model snap shot	69
Figure 4. 9 : Dataset snap shot for Naïve Bayes Model.....	70
Figure 4. 10 : Data structure snap shot for Naïve Bayes Model	71
Figure 4. 11 : Dataset partitioning for Naïve Bayes model	71
Figure 4. 12 : Model building and computation of accuracy for Naïve Bayes Model snap shot .	72
Figure 4. 13 : ROC Curve result snap shot for Naïve Bayes predictive model	74
Figure 4. 14 Importing dataset from csv file to RStudio for Logistic Regression snap shot.....	75

Figure 4. 15 : Dataset for Logistic Regression Model	76
Figure 4. 16 : Dataset structure for Logistic regression model.....	76
Figure 4. 17 : Dataset partitioning for Logistic regression snap shot.	77
Figure 4. 18 : Logistic Regression Model building and prediction accuracy snap shot.	77
Figure 4. 19 : The output of logistic regression model snap shot.....	78
Figure 4. 20 : Confusion Matrix for Logistic Regression snap shot.	79
Figure 4. 21: ROC Curve result snap shot	81
Figure 4. 22 : Required parameters for a sing vehicle.	84
Figure 4. 23 : Building a model	85
Figure 4. 24: Saving a mode to specific directory	85
Figure 4. 25 : Capturing vehicle’s important parameters	85
Figure 4. 26: Putting the captured data into appropriate model format.	85
Figure 4. 27: Loading the saved model to R studio	86
Figure 4. 28: Making prediction using the newly captured data.....	86
Figure 4. 29: Capturing vehicle’s information (required Parameters).....	86
Figure 4. 30: putting the captured data into appropriate model format.	87
Figure 4. 31: loading the saved model to R studio	87
Figure 4. 32: predictions using the newly captured data	87

List of Tables

Table 3. 1: Procedure Comparison of KDD, SEMMA and CRISP-DM:	34
Table 3. 2 : Comparison of Data Mining process modes.....	34
Table 3. 3: Confusion matrix.	37
Table 3. 4 : underwriting raw data	40
Table 3. 5 : Claim raw data.....	41
Table 3. 6 : Attributes selected for experiment.....	44
Table 3. 7 : Attributes Transformation	45
Table 3. 8 : Policy Code Transformation.....	45
Table 3. 9 : Claim History Transformation.....	46
Table 3. 10 : List of vehicle makes	46
Table 3. 11 : List of purposes the vehicles are user for.	47
Table 3. 12 : Group of Age of the vehicles.....	48
Table 3. 13 : Carrying Capacity of the Vehicles.....	48
Table 3. 14 : Premium to be calculated based on the risk ration	49
Table 3. 15 : Dataset in R Studio	49
Table 3. 16 : Numerical representation of purpose of the vehicles.	59
Table 3. 17 : Age group of the vehicles under experiment.....	60
Table 3. 18 : Numerical representation of CC.	61
Table 3. 19: Comparison of prediction models.....	83
Table 4. 1: Required parameters for a sing vehicle.	84

Abstract

The purpose of this study was to identify risk items with high claim ratio in order to take an appropriate measures during underwriting process to save profit making risk items under motor class of business. Even if the motor class of business takes the big portion of premium collection in Ethiopian insurance industry, most insurance companies indicate motor class of business as loss making line of business in their annual report. The main cause for this loss contribution is there are some risk items with high claim ratio which consumes a lot of the premium from the pool. Identifying those risk items from profit making risk item, helps a lot for an insurance company to maximize its profit. To tackle the problem of high claim cost in motor class of business, predictive data mining techniques has been employed using SVM, Naïve Bayes and Logistic Regression predictive models. The dataset used for the experiment in this study was collected from Awash insurance company specifically from underwriting and claim data tables of motor class of business. After cleaning irregularities and incomplete data in the dataset, a total of 52,831 records have been used to train the models in the ratio of 80:20. Among the used predictive models Naïve Bayes model outperformed the other two scoring 97.56% of accuracy and 98.7% precision. The challenging part of this study is lack of uniformity in conducting underwriting process. The underwriter may use either configured rate premium calculation which is similar throughout the company or flat rate which is specific to a branch and customer. This creates lack of uniformity throughout the company in terms of premium calculation. On the other hand most of records under configured premium calculation rate are complete and the values of attributes selected for this study are mandatory for the underwriter to be captured during underwriting process. Since predictive data mining techniques are aimed to identify patter of records in the dataset, only those risk item which have got insurance coverage with configured premium calculation rate in Awash insurance are included under this study. The predictive modes have been checked by new risk item as a prototype which is different from testing date and the outcome confirms the models are well trained and work correctly.

Key words: Predictive data mining, Awash insurance, motor class of business, Risk Items

CHAPTER ONE

1 INTRODUCTION

1.1 Background of the Study

Insurance industry is supporting the overall wellbeing of an economy and providing various benefits to the society (Weerasinghe and Wijegunasekar, 2016). The lack of efficient management of financial industry may affect the industry in general and insurance companies in particular but the whole economy may result in unfavorable consequence, indemnification, a basis for credit, stimulating savings and providing investment capital (Rani and Gobena, 2017). Moreover, it is undeniable that insurance companies are playing a crucial role in supporting the economy of a country with respect to different point of views such as risk transfer and intermediation (Sambasivam and Ayele, 2013). According to Sambasivam and Ayele (2013), well managed insurance industry channels funds and transfer risk from one economic unit to another economic unit. The lack of efficient management of financial industry may affect the industry in general and insurance companies in particular but the whole economy may result in unfavorable consequence for the economy of the nation such as institutional insolvencies (Naveed et al, 2011) as insurance industry is a big portion financial industry. Insurance is a mechanism (or a service) for the transfer to someone's called the insurer of certain risks of financial loss in exchange of the payment of an agreed fixed amount. The payment is due before the contingent claim is serviced by the insurer (Hailegebreal, 2016).

Nowadays, there are two general categories of insurance coverage throughout the world; life and Non-Life insurance service (Olayungbo, 2015). Under Non-Life insurance, Motor insurance is one of the main insurance services provided by insurers where several insurance companies were formed to specifically underwrite motor insurance business long back 1890's (Rani and Gobena 2017). After acquisition of license from National Bank of Ethiopia as per the insurance business Proclamation No. 86/ 1994, insurance companies in Ethiopia transact insurance business under "general insurance/Non-Life" and/or "long-term/Life" insurance (NBE, 2012).

Currently there are 17 primary insurance companies in Ethiopia, which are actively participating in insurance business. Almost all insurance companies engaged into provision of different class of business available Under Non-life insurance service which may differ slightly in category or nomenclature. Nevertheless Motor insurance, Burglary and House Breaking, Consequential loss, Engineering, Fidelity guarantee, Fire and Lightning, Group personal Accident, Horticulture, Inland Carrier Liability, Inland Transit, Marine, Money insurance, plate Glass, Political Violence and Terrorism, Product liability, Professional Indemnity, public Liability , workmen's compensation etc., are common class of businesses transacts in all insurance companies in Ethiopia. According to Rani and Gobena (2017), the number of motor vehicles in Ethiopia has increased tremendously from time to time. This can be observed when we look at the streets of Addis Ababa, Adama, Bahir Dar, Mekele, Hawassa and the like, the traffic jam confirms and highlights the ever increasing role of motor transport in the everyday life.

The Growth of motor vehicles in number is caused by a variety of factor such as increasing trend of urbanization, improvement of the standard of living, relocation of employees from their working place (most condominium houses are built far away from where they used to live.

Unlike other classes of business, motor insurance doesn't require proportional re-insurance arrangement (Rani and Gobena 2017). As a result, the premium collected from motor class of business remain in primary insurance company which motivate insurance companies for the wide spread of this class of business. Among Non-life insurance class of businesses, Motor class of business takes the lion share in terms of premium contribution and claim cost which is a common feature in most developing countries (Alhassan and Biekpe, 2015). For instance United Insurance S.C reported the premium contribution of Motor class of business in 2016/17 as 60% of the total portfolio and 85.7 loss ratio (claim ratio) while Awash Insurance S.C reported under the same reporting year premium contribution of motor class of business 65.5% and claim ration 87% (UIC and AIC annual report of 2016/17). Thus, a close analysis in this class of business makes insurance companies more profitable and positive contribution to the industry as a whole. Accordingly, there is a need to predict and classify risk items (the actual risk that the customer needs to be covered / insurable item, for instance, motor private familiar model, motor commercial own damage etc.) as loss makers and /or profit makers in the motor class of business using data mining techniques.

1.2 Motivation

As the middle class of society in most African countries expanding and their economy showing progress from time to time, motorization of society in both commercial and personal is increasing at increasing rate from time to time; and motor insurance has become the major line of insurance service to build market share through highly competitive pricing and it is the one that is steadily becoming compulsory across the continent (Africa Insurance Barometer 2017 Market Survey).

The motor class of business is the dominant premium generating line of business, accounting 41% of industry premiums compared to 59% to the rest of Non-Lifer class of business in 2012 in Africa (Alhassan and Biekpe, 2015). Motor class of business is not only a major source of premium collection in Africa in general and in Ethiopian insurance industry in particular, it is also the major source of claim costs which as a result leads this class of business a loss making class of business. For instance Awash Insurance S.C which is one of the leading insurance company in private insurance sector, in Ethiopia, reported in 2016/17 the premium contribution of motor class of business was 65.5% and claim ratio of the class was 87%. This shows how serious the claim cost of motor class of business is. But it doesn't mean that all risk items of motor class of business are loss makers; this can be evidenced by many customers who reported "No Claim Discount (NCD)" every year at policy renewal period. Therefore, there should be a mechanism to differentiate those risk items of motor class of business which are loss making from those which are profit making risk item of the class. This in turn helps the decision maker to take appropriate measure for prudent underwriting which can be premium adjustment or avoidance from the pool to increase the profitability of the company in particular and wellbeing of the insurance industry in general.

1.3 Statement of the problem

As any profit making industry, Insurance companies require earning profit to be sustainable in the current competitive environment (Berhe and Kaur, 2015). Motor insurance is the most prevalent and the largest sector in non-life insurance in most countries in the world (Ayele, 2014). According to Ethiopian insurance industry report in 2017/2018 total premium collected from motor class of business was Birr 4.4 billion which is 46% of the total Non-life insurance business.

However, motor class of business is not an attractive line of business in Ethiopian insurance industry and almost all insurance companies describe in their annual reports that motor insurance is consistently registered a negative results. Moreover, an investigation of major factors which determine insurance company's profitability has been conducted by many scholars on insurance companies' profitability and came up with different findings and conclusions. Among the many research works in the motor insurance, some of them have been given more attention in this study.

Yunos, Ali, Shamsyuddin, and Ismail (2016), made their research on prediction of motor claim frequency and severity using Artificial Neural Network (ANN) in their paper titled "Predictive Modeling for Motor Insurance Claims Using Artificial Neural Networks". The researchers investigated the capability of ANN as a potential technique to be applied in modeling the motor insurance claims problem. According to the authors, accurate predictive models in motor insurance claims are used for risk classification which is used as the formulation of different premiums scheme for the same coverage based on their category of the customers. The authors used Back Propagation Neural Network (BPNN) algorithm in ANN with a three-layer network structure of a back propagation (BP) learning algorithm to model the motor insurance claims. For the model evaluation, the researchers used 4 statistical methods, these are mean squared of error (MSE), root mean square of error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The result shows among the statistical methods MAPE error represent a smallest value of error for claim frequency and claim severity. Moreover, the authors described that BPNN model is successful in predictive modeling the Malaysian motor insurance claims by using several of network structures.

Shi, Feng and Ivantsova, (2015) in their research paper "Dependent frequency-severity modeling of insurance claims" claims that at individual level. Predictive models are used for risk classification and to determine the premium loadings for each policyholder at aggregate level. Moreover, it quantifies the risk of a portfolio or a block of business, which helps insurers choose the appropriate level of risk capital and treaty or facultative reinsurance arrangements. The authors used in their research conditional probability model and the mixed copula model to check whether there is dependency between severity and claim count (frequency). Moreover, the authors explained that mixed copula regression model contains two parts. The first part is The Poisson

regression component represents the number of claims in a group of policy holders, while the second part is the Gamma regression component models which corresponds average to the claim size of the group. According to the authors, mixed copula regression model simulation result shows that there is correlation between severity and frequency of insurance claim which implies the occurrence of one of them helps to predict the other.

Weerasinghe and Wijegunasekara (2016) conducted their research paper Namely “Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims”. According to the authors, Insurance industry plays a great role in the development and wellbeing of the country’s economy. Moreover, Insurance claims are a significant and costly problem for insurance companies in all sectors of the insurance industry. They got focused on factors that affect the number of claims so that identifying these factors helps to determine the amount of premium to be charged. But it is not easy as there are many factor contributing to the result. As a solution, the researchers proposed to build a predictive modeling in Motor insurance claims which is accurate classifier system of model in order to overcome the problem.

The researchers used three different data mining techniques in comparison to classifying factors which affect auto insurance claim. These are Neural Network, Decision Tree and Logistic Regression and the result shows ANN is the best predictor with 61.71% overall classifier accuracy. Decision tree came out to be the second with 57.05% accuracy and the logistic regression model indicated 52.39% accuracy.

Williams & Huang (2013), used Knowledge Discovery in Databases (KDD) process which uses decision tree techniques to identify significant areas of risk within an insurance portfolio in their paper titled “Knowledge Discovery in Database for Insurance Risk Assessment”. According to Williams & Huang (2013), isolating and understanding areas of risk is a significant task performed by an insurer. They said “Insurance is a business of risks” because the existence of Insurance is to prorate insurance risks across the pool of insured. The researchers claimed that “post data mining” analyses is significant which is expected to lead to a better understanding of insurance risk and to a finer tuning of insurance premium setting.

Hanafizadeh and Paydar (2013) in their research paper titled “A Data Mining Model for Risk Assessment and Customer Segmentation in the Insurance Industry” discussed in detail about customer categorization according to the claim amount they registered to the insurance company. According to the authors, a customer is taken as the essential factor in producing income in any business and increasing profitability of a company. Most insurance companies experience great loss as far as automobile insurance is concerned while most developed countries have attempted to increase the productivity and profitability of their insurance industry using the risk segmentation system (Hanafizadeh and Paydar, 2013). This is because lack of measures of determining the risk in automobile insurance leads to computing unfair rates. The authors segmented the customers based on the vehicle type that the customers drive which are pick-ups cars, Pride cars and Xantia cars and also gender of the drivers as segmentation parameter. They used K-means algorithm for segmentation process. The result of this research paper shows, pick-ups have higher risk than the other two makes and Pride cars have higher risk than Xantia cars. More over the data analysis indicated that male policyholders have higher risk than female ones.

Burri,Bojja and Buruga (2019), conducted their research paper titled “Insurance Claim Analysis Using Machine Learning Algorithms” and explained the Machine language as a data facilitator which to be converted to knowledge for decision making. According to Burri, Bojja and Buruga (2019) insurance companies are extremely interested in the prediction of the future. This is because accurate prediction gives them a chance to decrease financial loss for the company in connection with claim cost. Moreover, forecasting the upcoming claims helps to charge competitive premiums that are not too high and not too low. It also contributes to the improvement of the pricing models. This helps the insurance company to be one step ahead of its competitor. The authors used Naïve Bayes Updatable, Naïve Bayes, Multi-Layer Perceptron, J48, Random Tree, Logistic Model Tree (LMT), and Random Forest for this research paper and the result shows Logistic Model Tree (LMT), Random Forest algorithms have given better claim prediction when compared with the rest of classification algorithms.

Taking into account the content of these research papers, different issues related to insurance claim have been addressed. For instance, prediction of motor claim in relation with frequency and severity has been employed in detail. The researcher found out that there is a positive relationship between motor claim frequency and the severity of that claim. That means the higher the claim frequency the higher the severity of that claim. Then intention behind the research was formulation of different premiums scheme for the same coverage based on claim frequency of the customers. On the other hand comparison of predictive models for motor insurance claim has been also studied in detail. The researchers have tested different predictive models on the same data for the best accuracy level of prediction of claim costs to adjust the premium calculation scheme accordingly.

The analysis of policyholders claim information is the other issue that have been addressed. Categorization of policyholders according to the claim cost they brought to the pool was another focal point in the research papers. The purpose of the analysis was to allocate higher premium charge to customers who brought more claim cost to the company. Insurance Risk understanding was another very important factor that have been discussed in the research paper. It is also disclosed that understanding the risk that the insurance company assumed during under writing helps to adjust the premium amount to be collected from the insured.

From the analysis we can understand that these research papers gave more attention either to find efficient predictive models for insurance claim prediction or finding a way to categorize customers according to the claim amount that they brought to an insurance company within specific period of times. Contributing a lot in claim prediction as it is attempted to analyze, most of the studies did not address the risk item level claim cost analysis in their study. The importance of risk item level study of insurance claim prediction is saving other risk items which are profit makers.

Regarding to profitability of insurance industry in Ethiopian there are some research works conducted so far. Some of the authors who conducted insurance industry include

Mehari and Aemiro (2013), *“Determinants of insurance companies’ profitability Analysis of insurance sector in Ethiopia”*

Demise and Hailegebreal (2016), *“Determinants of insurance companies’ profitability Analysis of insurance sector in Ethiopia”*

Sambasivam and Ayele (2013), *“A Study on Performance of Insurance Companies in Ethiopia”*

Hailegebreal (2016). *“Macroeconomic and Firm Specific Determinants of Profitability of Insurance Industry in Ethiopia”*.

All these researches conducted on Ethiopian insurance industry mostly focused on assessment and identification of factors affecting the profitability of the industry rather than Risk item level claim cost analysis which is actually the main problem of the industry that should be addressed.

From the discussion made with claim managers of big branches of Awash insurance Company who have given a privilege to settle claims, the critical issue in motor class of business is identifying specific risk-items which eat the premium of others because of their high claim cost. The claim managers agreed that motor class of business is not a loss making business in general. Some risk categories from motor class of business are very profit making business and others are loss making business. So, the issue is how to identify this risk category or risk items from the others in the same class of business which is motor class of business? Moreover, senior underwriter and managers of big branches of the company (Awash insurance company), emphasized that isolation of risk items in the insurance industry is the most important issue in all Ethiopian insurance companies in general and for their branch offices in particular. According to managers, a study conducted at risk item level helps the industry to prepare their strategic plan accordingly whereas for branches especially in Awash insurance company, most benefits including salary increment and bonus for employees of all branches depends on the claim ratio they registered, which is directly related to the profitability of the branch. Therefore, this study applies data mining technique to identify insurance’s risk claim cost at risk level under motor class of business, more specifically to identify which risk items brings more claim cost to the pool.

1.4 Research Question

- ✓ What are the existing mechanisms and its gaps that insurance companies apply to identify the most loss making risk items in motor class of business?
- ✓ How to develop predictive model that predicts high risk items that brings more claim cost?
- ✓ How can predictive data mining employed to predict risk items with high claim cost?

1.5 Objective of the Study

1.5.1 General objective

The main objective of this study is to predict the most loss making risk item in Motor Class of Business using data mining technique by taking Awash Insurance Company S.C. as a case.

1.5.2 Specific objective

- To identify attributes that predict risk item with high claim cost
- To select appropriate algorithm to be used for building the model that can be applied to predict risk items with high claim cost
- To build data mining models on the preprocessed dataset.
- An activity after model building
- Evaluate the model performance

1.6 Scope of the Study

The scope of this study is bounded to Awash insurance company's motor class of business; giving more due attention to the claim they incur in the company since 2012. All the required data will be collected from Awash Insurance Company Database.

1.7 Limitations

As the operation activities of Awash insurance company is automated since November 2012 all the records process takes place through a system. There are three ways of capturing data in the company using the system. Configured rate, flat rate and fixed amount. For flat rate and fixed amount, some important attributes which are very important for this research are optional and not captured in most cases. Therefore, the data collected for this research paper is limited to policies which have been done using configured rate where most of the fields are mandatory and the user is enforced to capture them all.

1.8 Significance of the study

This study will contribute a lot for all insurance companies in general and Awash insurance company in particular. From information published on web sites of most insurance companies operating in Ethiopia, the type of insurance coverage they give are almost similar. Therefore any insurance company can refer the finding of this paper and make policy and procedure(rule of conducting a business) regarding the risks they assume during underwriting process. On the other hand the research result of this paper can reflect exactly what the risk level profitability of motor class of business looks like in Awash insurance company. Since the actual data is taken from Awash insurance company database, the outcome of this paper will be a fact that shows the contribution of each category of risk item in the company. Therefore, based on the outcome of this research, the company can make different decisions including loading premium so that the possible claim cost can be served or giving premium discount based on the customers' business context on hand.

CHAPTER TWO

2 REVIEW OF LITERATURES

2.1 Introduction

Under this chapter conceptual discussion on insurance, risk and data mining in relation with insurance business will be conducted and review of literature related to application of data mining techniques in insurance business in general and in claim risk prediction in particular will be given more attention.

2.1.1 Introduction to Insurance

Under this chapter conceptual discussion on insurance, risk and data mining in relation with insurance business will be conducted and review of literature related to application of data mining techniques in insurance business in general and in claim risk prediction in particular will be given more attention.

The term Insurance is defined by many scholars as a social device providing financial compensation for the loss of misfortune. It may be seen as a kind of contributed fund, into a pool which all who are insured will pay proportionate contributions called premium (Dinku, 2000). As far as an insured paid the expected premium, he/she will have the right to claim on the fund which is contributed from all insured's for any appropriate payment when unexpected loss occurred. Based on the above definition, one can conclude that insurance exists to avoid the adverse effect of risk in day-to-day activities.

From legal point of view, insurance is a contract whereby one party, the insurer, undertakes, for a premium or an assessment, to make a payment to another party, the policyholder or a third party, if an event that is the object of a risk occurs (Outreville, 1998).

According to Osterville (1998), insurance is often defined as a contract of indemnity. That means the insured is not to make any profit out of the insurance but should only be compensated to the extent of the financial loss. Moreover, insurance is a mechanism (or a service) for the transfer of risk of financial loss to someone else called the insurer in exchange of the payment of an agreed fixed amount known as premium. The insured should pay the premium before the contingent claim is serviced by the insurer (Nowadays this is not an issuer in Ethiopia insurance industry since the National Bank of Ethiopia declared “No premium No cover” policy in August 2012). From the insured's point of view, insurance is a transfer of risk whereas from the insurer's point of view, insurance as a "pooling" mechanism of a large number of exposure units or risks (Outreville, 1998).

2.2 Development of Insurance in brief

From historical development point of view, insurance has its own part in history. Some authors claim that marine insurance is considered to be the oldest known type of insurance. According to Abebe (2000), an activities similar to marine insurance was practiced at least 1000years before the Christian era while the present day form of marine insurance probably began around the eleventh or twelfth century and the early development of life insurance was closely linked with that of marine insurance. The industrial revolution in Europe demanded the development of fire and accident insurances and then followed by motor, engineering and aviation insurances (Abebe, 2000). In all cases, community based activity is taken as the origin of modern age insurance business which grew to a huge commercial sector covering a wide range of policies in insurance industry

2.3 Principles of Insurance

Insurance coverage has its own principles which are dedicated and identifiable only for insurance business. The main objective of every insurance contract is to give financial security or compensation and protection to the insured from any future uncertainties and for that to happen, insured must never ever try to misuse this safe financial cover (Akrani, 2011). According to Sibindi (2013), in addition to the common law of contract, the following rules in particular are considered to be the basic principles of insurance and these insurance principles are well respectful by both parties called the insured and the insurer all over the world (Akrani 2011)

These insurance principles are known as “Principle of Utmost Good Faith” which is the insurance contract must be signed by both parties (insurer and insured) in an absolute good faith or belief or trust, “Principle of Insurable Interest” which is to say a person has an insurable interest when the physical existence of the insured object gives him some gain but its non-existence will give him/her a loss, “Principle of Indemnity” which stands for an insurance Contract signed only for getting protection against unpredicted financial losses arising due to future uncertainties, “Principle of Contribution” means if the insured has purchased more than one policy on the same subject matter. According to this principle, the insured can claim the compensation only to the extent of actual loss either from all insurers or from any one insurer, “Principle of Subrogation”, is applicable when the insured is compensated for the losses due to damage to his insured property, then the ownership right of such property shifts to the insurer, “Principle of Loss Minimization “which states that the insured must take all possible measures and necessary steps to control and reduce the losses in such a scenario and “Principle of Causa Proximate” this is when a loss is caused by more than one causes, the proximate or the nearest or the closest cause should be taken into consideration to decide the liability of the insurer(Akrani, 2011).

In broad sense, the insurance market is based on two fundamental characteristic that is the transfer of exposure from a single party to a large group and the sharing of all losses by all those in the group. This implies that insurance relies heavily on the Law of Large Numbers which is known as pool management (Sibindi, 2013).

2.4 Ethiopian Insurance History in Brief

Accordant to Hailu (2007) the emergence of insurance business in Ethiopia was closely linked to foreign insurance companies. In addition, foreign companies operating in Ethiopia participated actively in the establishment of the first domestic insurance company. According to various sources, the emergence of modern insurance in Ethiopia is traced to the Bank of Abyssinia which was established in 1905 as the first Ethiopian Bank (Hailu, 2007).

Hailu (2007) further discussed that there were foreign owned insurance companies mostly acting as an agent to their parent company operating in Ethiopia prior to the year 1951. The Imperial Insurance Company of Ethiopia Ltd. was the first domestic insurance company, established in 1951. Since then, until 1960 one domestic and numerous foreign insurance companies represented by agents were operating insurance Business in Ethiopia.

Moreover, Insurance market in Ethiopia was not regulated until 1960. The first proclamation was enacted in 1970 as a result of which foreign companies were prohibited directly or indirectly for transacting insurance business in Ethiopia based on this, some companies converted to domestic companies in line with the requirement of the law (Hailmichael, 2011)

In accordance with proclamation of 1970, regulation number 383/1971 was issued by the Ministry of Commerce, Trade and Tourism on matters which help to create conducive insurance market. The controller of insurance has issued license for 15 domestic insurance companies, 36 agents, 7 brokers, 3 actuaries and 11 assessors in accordance with the provision of the proclamation immediately in the year after the issuance of the law (Hailu, 2007).

Four years after the enactment of the proclamation, the military government that came to power in 1974 put an end to all private entrepreneurship. Then all insurance companies operating were nationalized and from January 1, 1975 onwards the government took over the ownership and control of these companies & merged them into a single unit called Ethiopian Insurance Corporation. In the years following nationalization, Ethiopian Insurance Corporation (EIC) became the sole operator. After the military government overthrown in 1991, Ethiopian Insurance Corporation was restructured and re-established as a public enterprise.

The Insurance Proclamation number 86/ 1994 opened the opportunity for establishing domestic private insurers and mandated the National Bank of Ethiopia to oversee the activities of both the state-owned and private insurance companies.

Currently, there are 17 insurance companies in Ethiopia with so many branches across the country.

List of insurance companies in Ethiopia:

- Africa Insurance Company S.C
- Awash Insurance Company S.C
- Global Insurance Company S.C.
- Lion Insurance Company S.C
- NIB Insurance Company
- Nile Insurance Company S.C
- Nyala Insurance Company S.C
- The United Insurance S.C
- Ethiopian Insurance Corporation
- Abay Insurance Company
- Berhan Insurance S.C.
- National Insurance Company of Ethiopia S.C.
- Oromia Insurance Company S.C.
- Ethio-Life and General Insurance S.C.
- Tsehay Insurance S.C.
- Lucy Insurance S.C.
- Bunna Insurance S.C.

2.5 Motor Insurance in Ethiopia

The role of Motor Vehicles in socio-economic of the society takes undeniable remarkable position due to its universal nature nowadays. The technology the manufacturers use to produce the vehicles is increasingly sophisticated and the use of them is enhanced from time to time and these day's motor vehicles are directly or indirectly dominating every pattern of human life (Temesgen 2004).

According to Temesgen (2004), the number, type and use of vehicles in Ethiopia is increasing from year to year and as the result the frequency and severity of motor accidents increased as well and leave the loss burden on the society. Moreover, large number of vehicles destroyed and damaged accidentally every year. Damage and destruction to property is almost certain to happen every day due to motor accidents.

Loss of life, series and minor injuries are sad news to listen very frequently. In Ethiopia, motor insurance plays a crucial role in the alleviation of the financial burden of policyholders subsequently to accidents. Moreover, motor insurance constitutes a significant proportion of the overall business of the insurance industry (Temesgen 2004).

In Ethiopia, the number of deaths due to traffic accidents is reported to be amongst the highest in the world. According to the WHO, in 2013 the road crash fatality rate in Ethiopia was 4984.3 deaths per 100,000 vehicles per year, compared to 574 across sub-Saharan African countries. Moreover, the number of people injured or killed in one crash in Ethiopia is about 30 times higher than that in the US (Kussia, 2017). In general, the scale and the severity of the problem are increasing from time to time and adversely affecting the economy of the country in general and the livelihood of individuals in particular (Person, 2008). The motor vehicle accident doesn't affect only the life and economic status of an individual but also the profitability of the insurance industry in the country. Nowadays most vehicles have got insurance coverage in one way or the other. Therefore, the increase in motor vehicle accident will directly increase the claim ration of insurance companies. Since motor insurance takes the lion share of premium contribution in Ethiopian Insurance Sector (as reported yearly from all insurance companies of the country), no insurance company dare to avoid motor insurance class of business but there should be a way to minimize the its claim ratio to enhance the profitability of the industry in general and profitability of the class in particular.

2.6 Risk

2.6.1 Introduction to Risk

Risk is the foundation of insurance (François, 2016). That means whenever one talks about an insurance, there is always a risk in his/her mind for which the insurance is required. It is clear that the purpose of insurance is to minimize or mitigate the corresponding risk of every opportunity to be enjoyed. According to François (2016), risk refers to perils to which the individual is objectively exposed at any time or it is a condition in which there is a possibility of an adverse deviation from a desired outcome that is expected.

Regardless of the manner in which risk is defined, its existence affects the economic performance of society and therefore imposes constraints on the optimum allocation of resources and on the economic development of all nations. Individual as well as business decisions are not made under conditions of certainty. Although the idea of risk may be difficult to conceptualize, all economic units are taking decisions that they consider beneficial to them compared to the risk it may be susceptible to and these economic costs are the primary reason why financial institutions attempt to avoid risk or minimize its impact. On the other hand, is a source of gain to the society as a whole (François, 2016).

Risk is the common denominator of all decisions made by human being especially financial sectors. The objective in taking these decisions is not to avoid risk but to recognize its existence and ensure that compensation is adequate for the risks being borne (François, 2016). According to the author, the presence of uncertainty about future income or future return from an investment means that there is a risk to be evaluated which is bearing the risk. Investors, for example, must be compensated (receive a reward) for giving up present consumption and bearing the risk of the investment. Almost all financial decisions including insurance involve some sort of risk-return tradeoffs (François, 2016).

2.6.2 Risk and Insurance

According to Insurance Council of Australia (2019), in insurance terms, risk is the chance something harmful or unexpected could happen. This might involve the loss, theft, or damage of valuable property and belongings, or it may involve someone being injured. Insurers assess and price various risks to work out how much they would need to pay out if a policyholder suffered a loss for something covered by the policy. This helps the insurer determine the amount (premium) to charge for insurance. To be able to put a financial value on a risk, insurers calculate the probability that the insured item or property might be accidentally lost, stolen, damaged or destroyed, how often this might occur and how much it would cost to repair or replace.

In simple words risk is danger, peril, hazard, chance of loss, amount covered by insurance, person or object insured (Sisk, 2018). The risk is an event or happening which is not planned but eventually happens with financial consequences resulting in loss. In terms of insurance a risky proposal can on one hand bring higher profits but on the other hand with high potential of losses. As risk can never be certain or predictable, it needs a kind of management. The risk management is nothing but a method to pre-judge the risk that may come up sometime in future. It is not prediction but a process of reducing the risk to a minimum level (Sisk, 2018). Risk management involves a number of measures that are used to keep the risk at possible minimum level. In our day to day life also we take many steps to keep the risk at lower level for example most people do not keep valuables at home and rather prefer to keep them in a bank locker by paying certain locker rent to the bank.

Similarly risk of life, health or property is reduced by purchasing a proper insurance. All these actions of individual persons are done under fear of uncertainty and unpredictability of future. Likewise in business and commerce also an element of fear of loss always exists if the risk components are not managed properly. Risk is a fear of happening something adverse and in order to prevent such adverse happenings a plan has to be in place to overcome such adverse happenings which is called as risk management (Sisk, 2018).

Motor insurance risk is the claim cost that the company is liable to paid to the extent of the sum insured of the vehicle. For many reasons, the probability of happening of claim for motor class of business is very high compared to other class of businesses which is evidenced by annual claim report of almost all insurance companies in Ethiopia. According to Ethiopian insurance industry report in 2017/2018, the volume of premium collection is very high from motor class of business, compared to other class of businesses and therefore, there should be a way to minimize motor insurance claim risk to maximize the profit a company can make out of this class of business.

From insurance company's point of view, everything that the customers or insured brings to the insurance company to be insured are known as risk item. If the customer bring for instance one automobile and one factory to be insured, we say the customer brought to the insurance company two different risk items. Motor insurance risk is the claim cost that the company is liable to paid

to the extent of the sum insured of the vehicle. For many reasons, the probability of happening of claim for motor class of business is very high compared to other class of businesses which is evidenced by annual claim report of almost all insurance companies in Ethiopia. Accordingly throughout this paper the risk item refers anything brought to the insurance company to be insured.

2.7 Data Mining

2.7.1 Introduction to Data Mining

Nowadays, knowledge is becoming a fundamental organizational resource that provides competitive advantage to compete in the industry they work in many organizations have collected and stored vast amount of data. However, they are unable to discover valuable information hidden in the data by transforming these data into valuable and useful knowledge Silwattananusarn and Tuamsuk (2012).Therefore, Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data; but managing knowledge resources can be a challenge and therefore many organizations are employing information technology in knowledge management (KM) to aid creation, sharing, integration, and distribution of knowledge throughout the target organization(Silwattananusarn&Tuamsuk, 2012).Knowledge management is a process of data usage and the basis of data mining is a process of using tools to extract useful knowledge from large datasets; data mining is an essential part of knowledge management (Dawei, J. 2011).

Knowledge management process focuses on knowledge flows and the process of creation, sharing, and distributing knowledge. Each of knowledge units of capture and creation, sharing and dissemination, and acquisition and application can be facilitated by information technology.

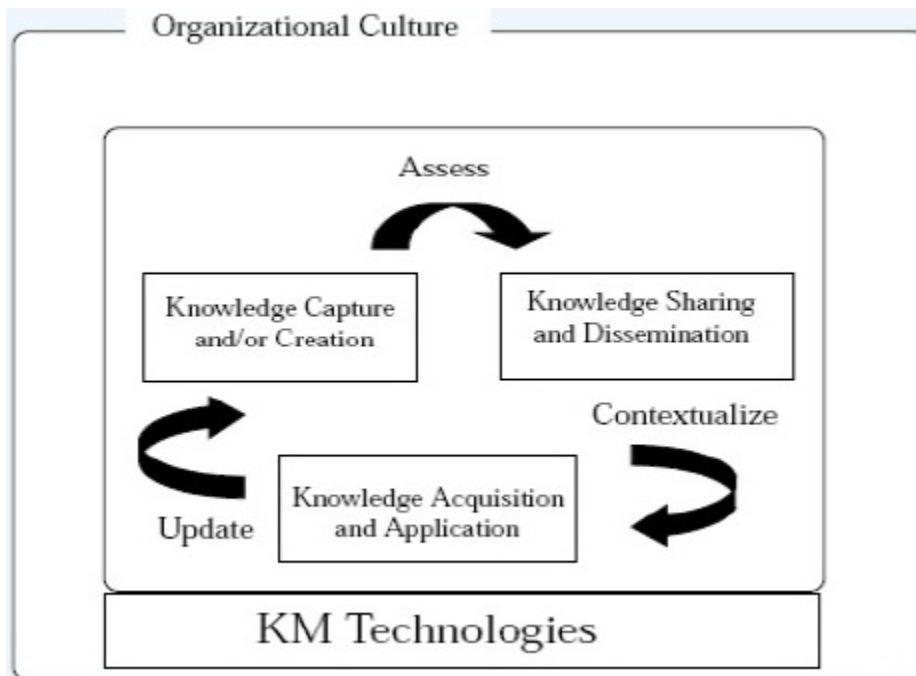


Figure 2. 1: KM Technologies Integrated KM Cycle (Source from Dalkir, K., 2005).

Data mining (DM) is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses (Deshpande and Thakare, 2010).

Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions and the automated, prospective analyses offered by data mining move beyond the analyses of past events provided both familiar tools like decision support systems. Moreover, data mining tools can answer the questions that traditionally were too time-consuming to resolve. But in case of DM preparation of databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations (Deshpande and Thakare, 2010).

According to Deshpande and Thakare (2010), Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. In earlier days data analysis process was manual and tough because domain knowledge was needed and understanding of statistical techniques was also needed. This manual process could not be applicable while facing the rapidly growing sizes and large dimensions of the data. A community of researchers introduced the term

named “data mining” to solve automating data analysis problem and discover the implicit information from the huge amount of data (Giordano, 1995)

Data mining is a step in (KDD) and aims to discover useful information from huge amount of data (Sumathi, Kannan and Nagarajan, 2016). The major role of data mining is applying various procedures and algorithms in order to retrieve patterns from huge amount of data and nowadays data can be taken from different kind of large volume of datasets in various formats like flat files, videos, records, texts, images, audios, scientific data and new kind of data formats. The data collected from different sources require proper data analysis for efficient decision making process (Sumathi, Kannan and Nagarajan, 2016).

Data mining is an interdisciplinary field of astronomy, business, and computer science, economic and others to discover new patterns from large datasets. Data mining technology can help the insurance firms for taking crucial business decisions (Umamaheswari and Janakiraman, 2014). The insurance sector is primarily dependent on customer’s base information and the most scenario of any insurance firm, effective management of customer data is one of the crucial one. With the help of data mining techniques, the customer data handled effectively help the market specialists and underwriters for decision making process in case of insurance business. Companies in the insurance industry collect huge amounts of data about their customers. In that situation, Data mining is very helpful for the firm to access the data and make analysis for prediction of different events or situations (Umamaheswari and Janakiraman, 2014)

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set (Singh&Kumar, 2012). These tools can include statistical models, mathematical algorithm and machine learning methods. Data mining often can improve existing models by finding additional, important variables, identifying interaction terms and detecting nonlinear relationships. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs. Specifically, data mining can help insurance firms in business practices (Devale and Kulkarni, 2012) by acquiring new customers, Retaining existing customers by performing sophisticated classification and analyzing policy selection.

2.7.2 Acquiring New Customers

One of the major problem in business process is the acquisition of new customers. Although the traditional approaches involve attempts to increase the number of customer by simply expanding the efforts of the sales department, it is less efficient in profit maximization as risk selection is not taken into consideration. A traditional sales approach is to increase the number of policyholders by simply targeting those who meet certain policy conditions. That means much of the marketing effort may yield little return. At some point, sales become more difficult and greater marketing budgets lead to lower and lower return. Hence it is important to identify population segments among already insured customers and others using data mining techniques.

2.7.3 Retaining Existing Customers

As acquisition costs increase, insurance companies are beginning to place a greater emphasis on customer retention programs. Experience shows that a customer holding two policies with the same company is much more likely to renew than a customer holding a single policy. Similarly, a customer holding three policies is less likely to switch than a customer holding less than three. By offering quantity discounts and selling bundled packages to customers, such as home and auto policies, a firm adds value and thereby increases customer loyalty, reducing the likelihood the customer will switch to a rival firm.

2.7.4 Classification: Databases Segmentation

To improve predictive accuracy, databases can be segmented into more homogeneous groups. Then the data of each group can be explored, analyzed and modeled. Depending on the business question, segmentation can be done using variables associated with risk factors, profits or behaviors. Classification algorithms requires classes which are defined based on data attribute values. They often describe these classes by looking at the characteristics of data already known which belong to the classes. As a result, insurance companies can more accurately predict the likelihood of a policy based on some attributes like premium amount, claim amount, Risk Item, etc.

2.7.5 Data Mining Techniques

Data mining techniques have been applied to various insurance domains to improve decision making. Data mining use predictive modeling, market segmentation, market basket analysis to answer business questions with greater accuracy. Various data mining techniques are used for the insurance industry development including Classification, Clustering, Regression and Association rules (Jayanthi Ranjan, 2009)

2.7.5.1 Classification

The types of classification under data mining are supervised classification and unsupervised classification. Various classifiers are used for the classification algorithms such as Decision tree, Bayesian classifier, neural network, Support Vector Machine etc. In classification, class attribute values are discrete. Given a set of data elements, classification maps each data element to one of a set of pre-determined classes based on the difference among data elements belonging to different classes. The goal is to discover rules that define whether an item belongs to a particular subset or class of data (R Joseph, Hlmani and Letsholo , 2016). And therefore, the customer database can be segmented into homogeneous groups and data mining classification algorithms can be applied on the dataset to organize them into intended classes.

2.7.5.2 Clustering

Similar to classification, clustering is the organization of data in classes. However, they are different in that unlike classification, in clustering, class labels are unknown and it is the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity). In general clustering is used for identification of similar classes of objects. It's used for grouping based on for example customer's behavior. It is applicable for customer segmentation and targeted marketing. Types of Clustering are partitioning methods, Hierarchical agglomerative methods, Density based methods Grid based methods and Model based methods

2.7.5.3 Regression

In regression, class attribute values are real numbers. For instance, if we wish to predict the stock market value (class attribute) of a company given information about the company (features). The stock market value is continuous; therefore, regression must be used to predict it. The input to the regression method is a dataset where attributes are represented using x_1, x_2, \dots, x_m (also known as regressors) (R Joseph, Hlmani and Letsholo , 2016). The class attribute is represented using Y (also known as the dependent variable), where the class attribute is a real number and the relation between Y and the vector $X = (x_1, x_2 \dots x_m)$ (R Joseph, Hlmani and Letsholo , 2016). Regression analysis is used to model the relationship between one or more independent and dependent variables. In insurance firm, more complex techniques are needed to predict future values. Some of regression types are: Linear regression, Non-linear regression, Multi-variate linear regression and Multi-variate non-linear regression

2.7.5.4 Prediction

Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply predicted. For instance, sales volumes, stock prices, product failure rates and insurance premium calculation rates are all too tough to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (logistic regression, decision trees, or neural nets) may be necessary to forecast future values (R Joseph, Hlmani and Letsholo , 2016).

2.7.5.5 Association

Insurance companies faced a lot of problems on customer retention now a days. Association is used for this task, because it finds all the association where customers bought a frequent item set. Association helps business firms to make certain decisions. Market basket analysis and cross selling programs are typical examples for which association modeling is usually adopted. When the customers want to insure some policies, then this technique helps us finding the associations between different items customer preference.

2.7.5.6 *Neural networks*

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or ambiguous data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques (R Joseph, Hlomani and Letsholo, 2016). These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs (R Joseph, Hlomani and Letsholo, 2016).

2.8 **Predictive models**

For this study three predictive models known as SVM, Naïve Bayes and Logistic Regression are used SVM which is generally capable of delivering higher performance for small and medium dataset in terms of classification accuracy (Srivastava and Bhambhu, 2010). According to the author, SVM is generally are capable of delivering higher performance for small and medium dataset (in our case 52, 831 records) in terms of classification accuracy. SVMs can learn a larger set of patterns and be able to scale better (R Joseph, Hlomani and Letsholo, 2016). Moreover, SVM has the ability to update the training patterns dynamically whenever there is a new pattern during classification and able to model complex nonlinear decision boundaries and are less prone to over fitting than other methods.

Naïve Bayes algorithm is very easy to use and effective if the training set is large enough. It means as the training set gets larger, and the results get more and more accurate (R Joseph, Hlomani and Letsholo, 2016). According to the authors, despite its simplicity, Naive Bayes can often out perform more sophisticated classification methods and the classifier will converge quicker than discriminative models like logistic regression, so it needs less training data. Naïve Bayesian classifiers simplify the computations and exhibit high accuracy and speed when applied to large databases. Moreover, Naïve Bayes algorithm ha an ability of handle noisy data, continuous and discrete data and make probabilistic prediction.

Logistic Regression is intrinsically simple, it has low variance and so is less prone to over-fitting and it is faster and more reliable when the dimension gets large. Moreover, Logistic Regression can easily update the model to take in new data (using an online gradient descent method) (R Joseph, Hlomani and Letsholo,2016).

CHAPTER THREE

3 RESEARCH METHODOLOGY

3.1 Introduction

Research methodology explains what research methods are going to be used, the choice of the research design and a strategy of data collection, management and analysis (Mahajan, 2010).

This chapter deals with Methodology and design of the experiment where methods, techniques, tools and algorithms applied to address research questions discussed. In general, understanding research methodology is systematic, theoretical analysis of the methods applied to a field of study. It includes the theoretical analysis of the body of methods and principles associated with a branch of knowledge. Methodology is not a solution for a particular issue, but it offers the theoretical understanding of which method or best practices will be applied to specific case (Irny and Rose, 2005).

3.2 The Data Mining Process

In general data mining task can be classified into two categories: Descriptive mining and predictive mining (Yan & Xie 2009). Descriptive mining is the process of drawing the essential characteristics or general properties of the data in the database. Clustering and association rules are example of descriptive mining. Predictive mining is the process of inferring patterns from data to make predictions. Predictive mining techniques involves task like classification and time series analysis.

The life cycle of a data mining project consists of six phases (Larose, 2005). According to Deshpande & Thakare (2010), the sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

3.2.1 Understanding of the problem

This is a preliminary phase that highlights the understanding of the objectives of data analysis and the converting of these requirements and the problem formulated into a definition of DM problem (Olegas Niaksu, 2015). In this phase it is determined the initial plan of achievement of goals,

defining the success criteria and basic analysis about the current problem and its solution visualizing the future research goal of the project which will be translated into Data mining goal later on.

3.2.2 Understanding of the data

This phase starts with the gathering of initial data. The problems of data quality must be identified and are created the initial assumptions which datasets can be of interest for further steps. (Olegas Niaksu, 2015). Therefore, at this step data is collected and decision about relevance of date, completeness, redundancy, format and size etc. will be checked according to Data Mining goal.

3.2.3 Preparation of the data:

The major activities of the data preparation phase heavily depend on the features and the quality of the original raw data. Some of the basic tasks of data preparation involve the selection of attributes, data transformation, classification, normalization, data cleaning and checking data completeness (Olegas Niaksu, 2015). And therefore, this is the step at which the researcher decides which data is useful and which one is not according to the Data Mining goal. Finally the data that meet the specific input requirements for the Data Mining tools selected.

3.2.4 Data modeling:

In this phase, a suitable selection of modeling techniques, algorithms, or combinations of them will be done. Then, optimal algorithm parameters' values are chosen (Deshpande and Thakare, 2010). Therefore, this the step at which the process of extracting knowledge from processed data is done.

3.2.5 Evaluation of the discovered knowledge:

In this stage the model selected at step (d) above is thoroughly evaluated and reviewed. At the end of this phase, a decision on the use of the data mining results should be reached (Deshpande and Thakare, 2010). This is a process of checking whether the discovered knowledge is novel, important and understandable. As per the result obtained at step 4 evaluation will be done whether the obtained result shows real problem of the Awash Insurance Company.

3.2.6 Use of the discovered knowledge.

The purpose of the whole process from (a) to (e) discussed above is to increase the knowledge of the data, the Knowledge gained will need to be organized and presented in a way that the customer cause it (Deshpande and Thakare, 2010). Moreover, when and how to use the discovered knowledge is part of this phase.

3.3 Data Mining Process Modeling

Currently there are many Data Mining process models available for data mining. Some of them are *Knowledge Discovery in Database (KDD)*, Sample, Explore, Modify, Model, and Assess (SEMMA), *Cross-industry standard process for data mining (CRISP-DM)* etc. are well known Data Process models

3.3.1 Knowledge Discovery Databases (KDD)

Knowledge Discovery Databases (KDD) is the process of extracting the hidden knowledge from databases. KDD requires relevant prior knowledge and brief understanding of application domain and goals. KDD process model is iterative and interactive in nature. KDD involves five different steps or stages (Xiao Zhu, 2017).

These are: Selection, Preprocessing, Transformation, Data Mining and Interpretation / Evaluation

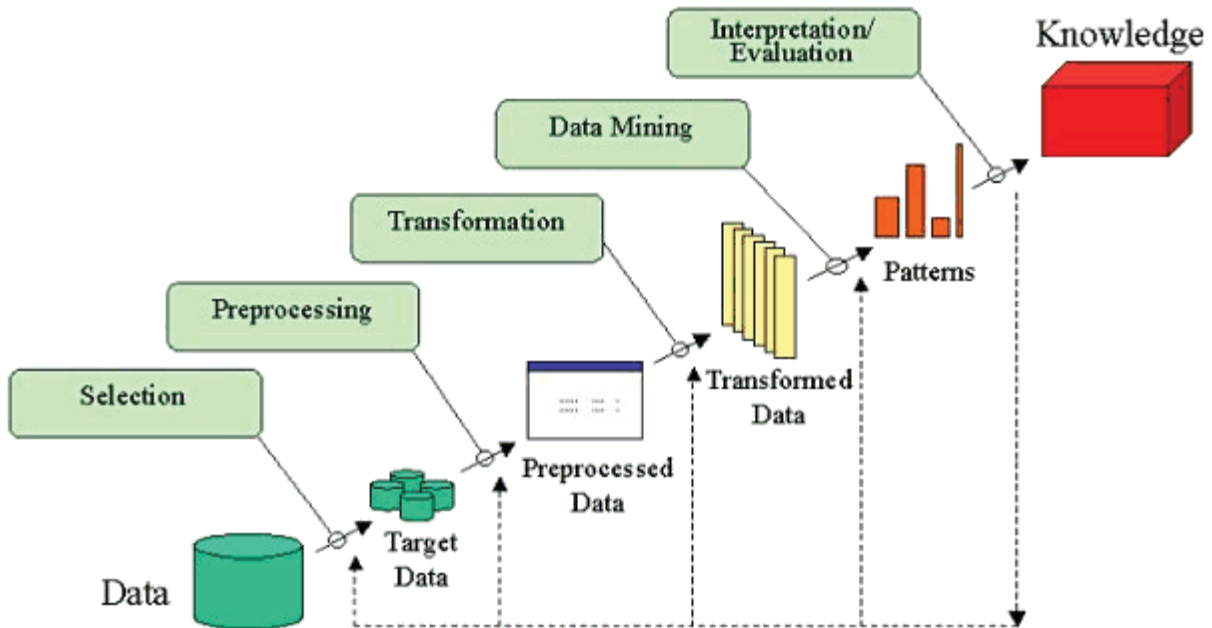


Figure 3. 1 : The Five Stages of KDD, Source: Xiao Zhu, 2017

3.3.2 Sample, Explore, Modify, Model, and Access (SEMMA)

SEMMA is data mining method developed by Institute for Advanced Statistics (SAS). It offers and allows understanding, organization, development and maintenance of data mining projects. It helps in providing the solutions for business problems and goals. SEMMA is linked to SAS enterprise miner and basically logical organization of the functional tools for them. It has a cycle of five stages or steps. These are: Sample, Explore, Modify, Model and Access.

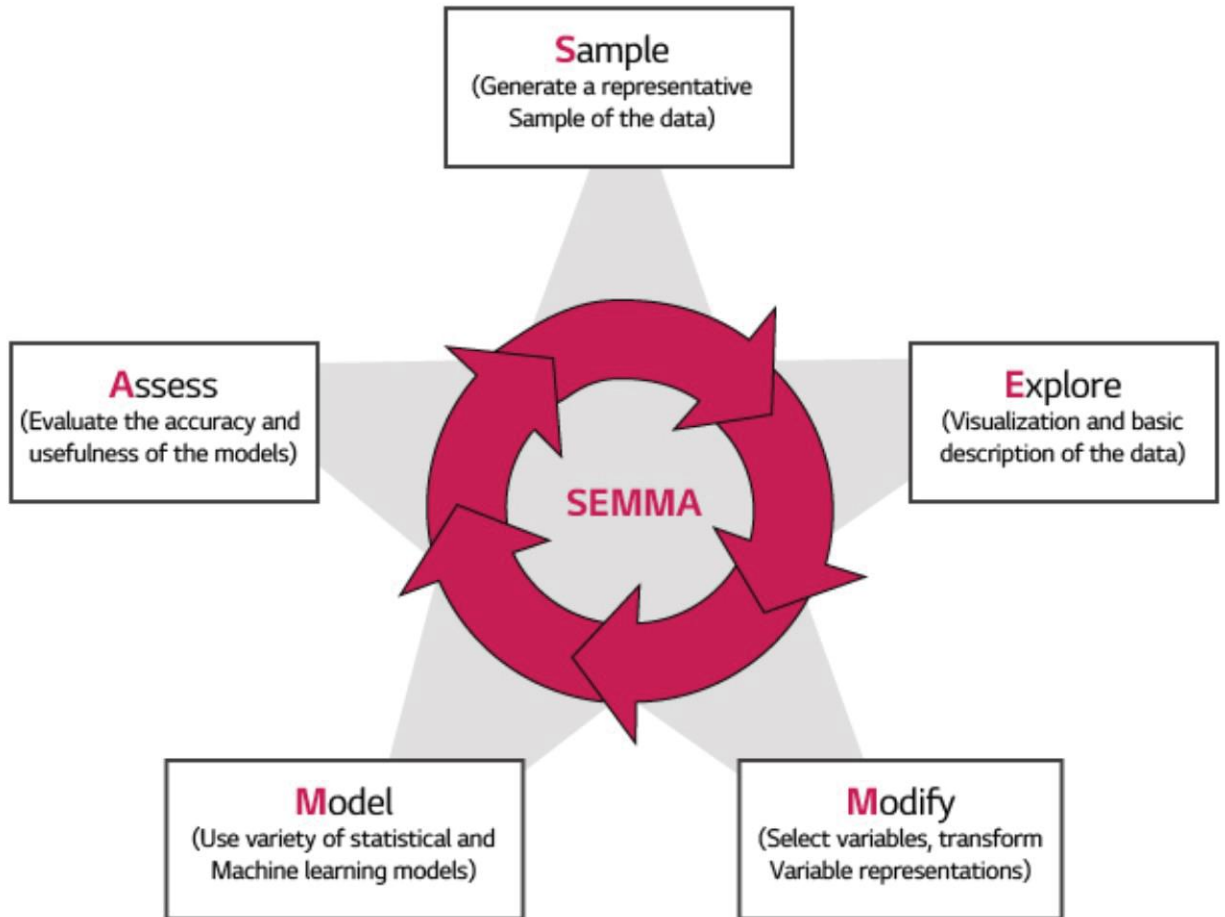


Figure 3. 2: SEMMA Cycle. Source: Xiao Zhu, 2017

3.3.3 Cross-Industry Standard Process for Data Mining (CRISP-DM) processing model

Cross-Industry Standard Process for Data Mining (CRISP-DM) was developed by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR in 1999, CRISP-DM 1.0 version was published and is complete and documented. It provides a uniform framework and guidelines for data miners (Shafique&Qaiser, 2014). It consists of six phases or stages which are well structured and defined (Shearer, 2000). These phases are described below.

- **Business Understanding:** This is the first phase of CRISP-DM process which focuses on and uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms.
- **Data Understanding:** This is the second phase of CRISP-DM process which focuses on data collection, checking quality and exploring of data to get insight of data to form hypotheses for hidden information.
- **Data Preparation:** This is the third phase of CRISP-DM process which focuses on selection and preparation of final data set. This phase may include many tasks records, table and attributes selection as well as cleaning and transformation of data.
- **Modeling:** This is the fourth phase of CRISP-DM process selection and application of various modeling techniques. Different parameters are set and different models are built for same data mining problem.
- **Evaluation:** This is the fifth stage of CRISP-DM process which focuses on evaluation of obtained models and deciding of how to use the results. Interpretation of the model depends upon the algorithm and models can be evaluated to review whether achieves the objectives properly or not.
- **Deployment:** This is the sixth and final phase of CRISP-DM process focuses on determining the use of obtain knowledge and results. This phase also focuses on organizing, reporting and presenting the gained knowledge when needed.

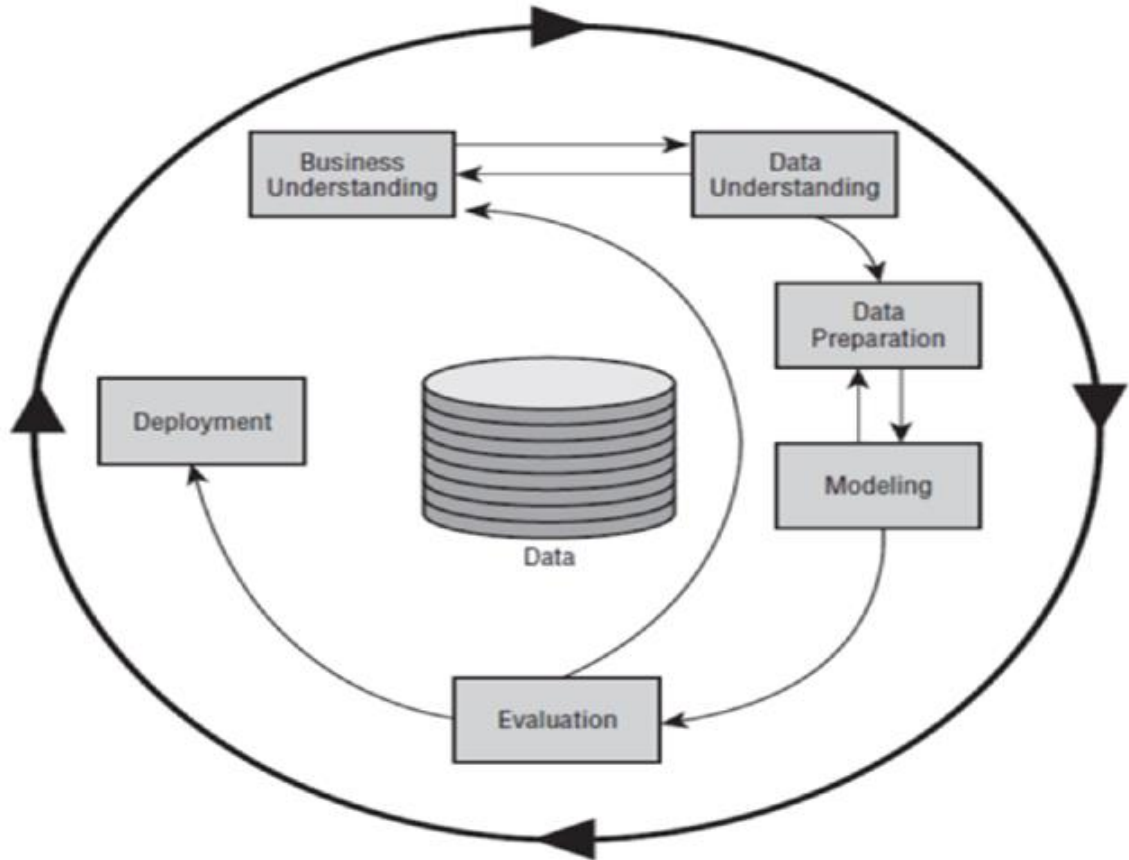


Figure 3. 3 : CRISP-DM Process Model, Source: Shafique, &Qaiser, 2017.

3.3.4 Comparison of KDD, SEMMA and CRISP-DM

According to Xiao Zhu (2017) the procedures conducting same function in the three data mining processes, both KDD and CRISP-DM, contain all procedure in SEMMA. Meanwhile, business understanding in CRISP-DM has an identical purpose with Pre-KDD procedures in KDD process, while Deployment in CRISP-DM summarizes the Post-KDD procedures. In the integrality view, CRISP-DM and KDD process is more comprehensive than SEMMA.

Table 3. 1: Procedure Comparison of KDD, SEMMA and CRISP-DM:

CRISP-DM	SEMMA	KDD
Business Understanding	N/A	N/A
Data Understanding	Sample	Selection
	Explore	Pre Processing
Data Preparation	Modify	Transformation
Modeling	Model	Data Mining
Evaluation	Assessment	Interpretation/Evaluation
Deployment	N/A	N/A

KDnuggets, a leading website of data mining, has conducted two polls about the usage of different data mining processes on 2007 and 2014. Among the 200 answers from poll 2007, 42 percent used CRISP-DM, 19 percent used their own model, 13 percent used SEMMA, 7.3 percent used KDD Process, 5.3 percent used their organization's model and 4 percent used some other model or no model. In the poll 2014, 43 percent of the 200 respondents used CRISP-DM, 27.5 percent used their model, 8.5 percent used SEMMA, 8 percent used other models, 7.5 percent used KDD Process and 3.5 percent used their organization's specific model. The results demonstrate CRISP-DM model's advantage on its overwhelmingly wide-spread usage. This is summarized in the following table

Table 3. 2 : Comparison of Data Mining process modes.

Model	Year	
	2007	2014
CRISP	42	43
Own Model	19	27.5
SEMMA	13	8.3
KDD	7.3	7.5
Org.Model	5.3	3.5
Other	4	8

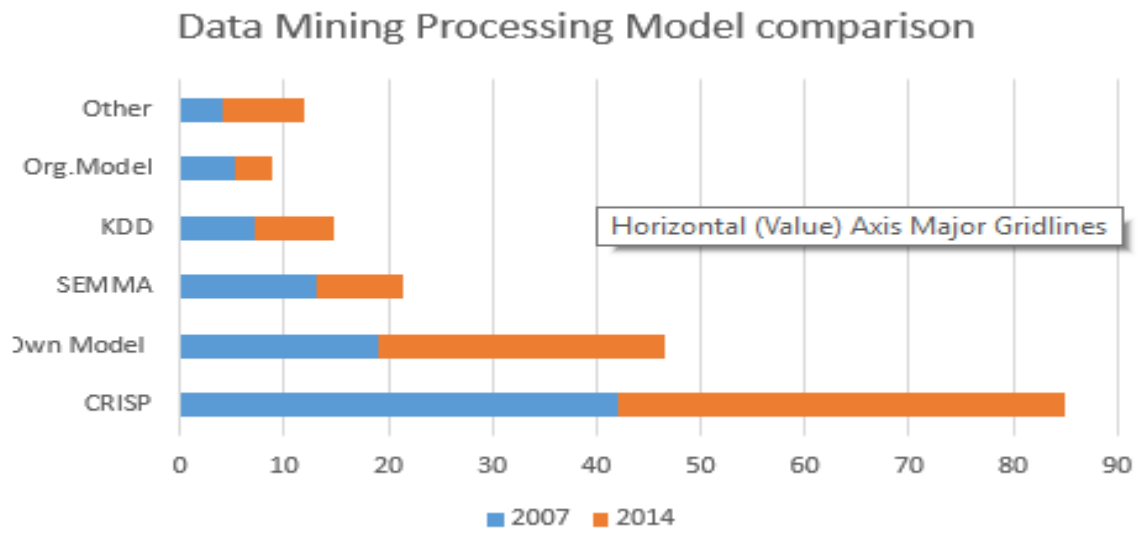


Figure 3. 4 : Comparison of Data Mining process modes

CRISP-DM has been consistently the most commonly used methodology as per KDnuggets polls starting in 2002 up through the most recent 2014 poll.

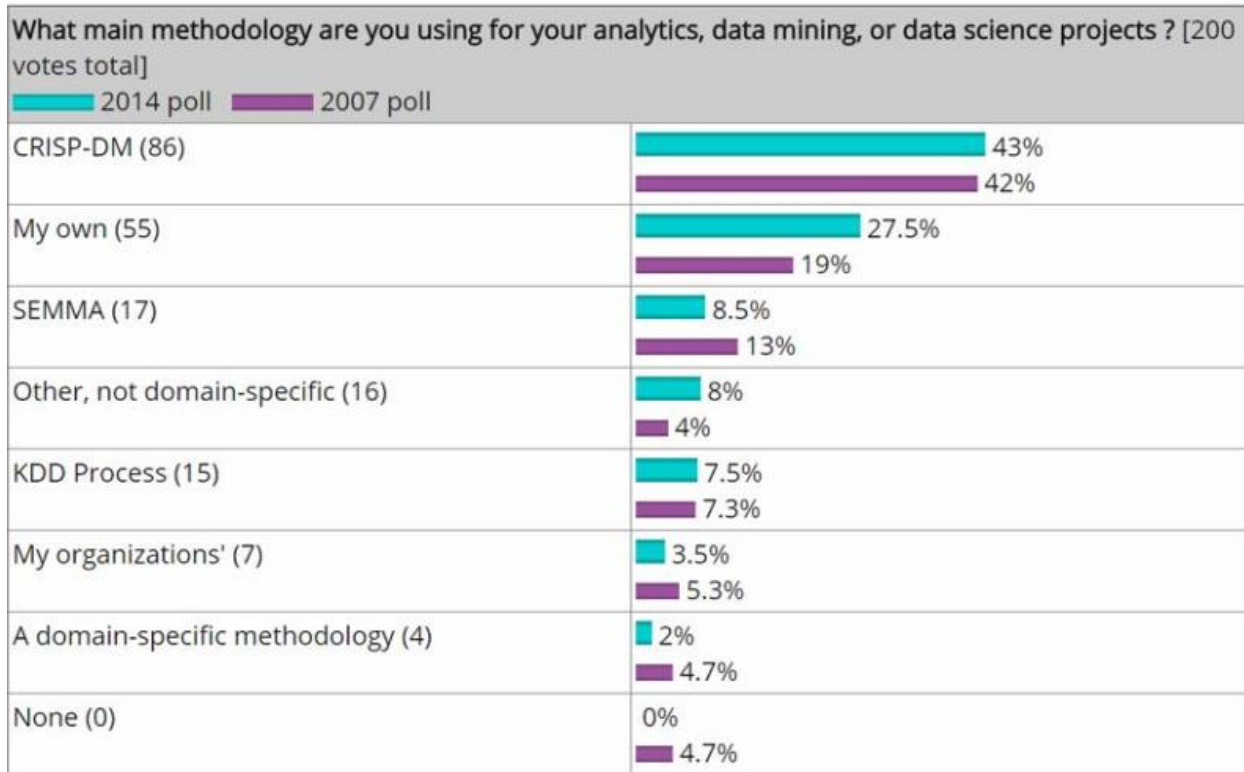


Figure 3. 5: KDnuggets poll results (Piatesky, 2014), Source: Xiao Zhu, 2017

From all the above information we gathered, CRISP-DM is the most popular Data mining process model. This paper is also employed CRIST-DM process model for the experiment which is going to be tested. For this study R studio and R programming user for the experiment conducted. R has characters like open source (One of the main reasons the adoption of R is spreading is its open source nature. R binary code is available for everyone to download, modify, and share back again), Plugin ready (There is a base version of R, containing a group of default packages that are delivered along with the standard version of the software. The functionalities available through the base version are mainly related to file system manipulation, statistical analysis, and data Visualization. And data visualization friendly (R complies with principles and techniques employable to effectively display the information and messages contained within a set of data)

3.4 Evaluation

At the fifth stage of CRISP-DM process model (or models) that appears to have high quality from a data analysis perspective will be built. According to Sastry and Babu (2013), before proceeding to final deployment of the model, it is important to thoroughly evaluate it and be certain that the model properly achieves the business objectives. Therefore, the predicted value will be evaluated against the actual values using Confusion Matrix. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix (Santra and Christy, 2012).

Table 3. 3: Confusion matrix.

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

The entries in the confusion matrix have the following meaning throughout this paper

- a. is the number of correct predictions that an instance is negative,
- b. is the number of incorrect predictions that an instance is positive,
- c. is the number of incorrect of predictions that an instance negative, and
- d. is the number of correct predictions that an instance is positive.

The other points related to confusion matrix are:

The accuracy (AC) is the proportion of the total number of predictions that were correct (Santra and Christy, 2012).

$$AC = \frac{a + d}{a + b + c + d} \quad (a)$$

True positive rate (TP) is the proportion of positive cases that were correctly identified (Santra and Christy, 2012).

$$TP = \frac{d}{c + d} \quad (b)$$

False positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive (Santra and Christy, 2012).

$$FP = \frac{b}{a + b} \quad (c)$$

True negative rate (TN) is defined as the proportion of negatives cases that were classified correctly (Santra and Christy, 2012).

$$TN = \frac{a}{a + b} \quad (d)$$

False negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative (Santra and Christy, 2012).

$$FN = \frac{c}{c + d} \quad (e)$$

Precision (P) is the proportion of the predicted positive cases that were correct (Santra and Christy, 2012)

$$P = \frac{d}{b + d} \quad (f)$$

3.5 Research design

The purpose of this study is to uncover the hidden risk item record patterns in the database of Awash Insurance S.C. As it is discussed under data Mining Process Modeling in detail, there are different data mining process models available among which Cross-Industry Standard Process for Data Mining (CRISP) is widely user process model (Olegas Niaksu, 2015). Taking into account the factors mentioned so far, under this study, the researcher also preferred Cross-Industry Standard Process for Data Mining (CRISP) and R programming for the experiment to be conducted.

3.5.1 Data collection

Under this study all the data was collected from Awash Insurance S.C. database. The data is limited to Motor commercial and motor private only from motor class of business. That means motor compulsory third party is not included in the dataset. Since motor third party insurance coverage is enforced by the state government, most of the crucial parameters like purpose of the vehicle, Age of the vehicle and carrying capacity of the vehicle are not available in the database; because they are not rating parameters for premium calculation. Moreover policies issued using flat rate calculation is excluded from the experiment. Flat rate is a rate defined for specific policy to determine the intended premium. This is used mostly for influential or major customers to strengthen the relationship they have with the company. All the policies included in this experiment are configured rate policies. The rate for configured policies is defined by the company executive management with consent of board of directors and included into the rate chart of the company so as to be distributed to all branches as a working rate chart which is similar throughout the company's operational units. In case of configured rate policies, the system (General Insurance Information System (GIIS)) automatically takes the integrated rate and calculated a premium. As a principle all the branches are expected to use configured rate in their underwriting activities which is also applicable for claim calculation. Therefore, this study intended to stick to the standard working rate which is configured rate. Moreover, for configured rate policies, all the parameters used for this experiment are compulsory so that the user is enforced to feed into the system. So they are available in the database for the experiment. But it doesn't mean all the data fed into the

system is correct. There were many records found incorrect data fed into the system just to pass the mandatory requirement of the system. All these misleading records are either corrected or excluded from the experiment. All policies with configured issued from 2012-2019 are included in this study. The record contains dependent variable representing either “Standard” for risks whose risk claim ratio is below the industry standard claim ratio which is Below 66% or “Loading” for risks whose risk ratio is above industry standard. Risk ratio (claim amount/premium collected) is calculated for a given risk items or customers.

This research paper also considered SVM which is generally capable of delivering higher performance for small and medium dataset in terms of classification accuracy (Srivastava and Bhambhu, 2010), Naïve Bayes which is effective and powerful algorithm for predictive modeling (Huang and Lei, 2011 and logistic regression models which preferable model for Categorical variable types and R programming will be employed as the building tool.

Table 3. 4 : underwriting raw data

A	B	C	D	E	F	G	H	I	J
Policy Number	Premium Amount	Risk Name	Vehicle Make	Purpose	Age of the Vehicle	Cylinder Capacity			
AIC/ADD/MTCM/011572/13	20115.76	Pickup over 2350cc		OWN GOODS	3	2494			
AIC/ADD/MTCM/011640/13	20054.05	Pickup over 2350cc		OWNGOODS	1	2500			
AIC/ADD/MTCM/011640/13	20054.05	Pickup over 2350cc		OWN GOODS	1	2500			
AIC/ADD/MTCM/011964/13	8372.48	Pickup over 2350cc	TOYOTA	OWN GOODS	10	4164			
AIC/ADD/MTCM/012321/13	43846.83	General Cartage - Trucks With Trailers	SCANIA	GENERAL CARTAGE	12				
AIC/ADD/MTCM/012472/13	13757.44	General Cartage - Tipper Nissan		GENERAL CARTAGE	11	0			
AIC/ADD/MTCM/012619/13	19055.7	General Cartage - Tipper Nissan		GENERAL CARTAGE	7	12503			
AIC/ADD/MTCM/012941/13	17909.45	General Cartage - Tipper Nissan		GENERAL CARTAGE	7	12503			
AIC/ADD/MTCM/013600/13	37452.24	General Cartage - Tipper Familiar model		GENERAL CARTAGE	1	9726			
AIC/ADD/MTCM/013695/13	27454	General Cartage - Isuzu Trucks - NPR		GENERAL CARTAGE	3	4570			
AIC/ADD/MTCM/013880/13	12840	Public Transport Isuzu Buses		PUBLIC TRANSPRT	7	4334			
AIC/ADD/MTCM/013898/13	14226.21	General Cartage - Trucks With Trailers	FIAT	GENERAL CARTAGE	24	13798			
AIC/ADD/MTCM/013920/13	5867.6	Pick up upto 2350cc		OWN GOODS	21	1595			
AIC/ADD/MTCM/014366/13	22263.12	Pickup over 2350cc		OWNGOODS	2	2494			
AIC/ADD/MTCM/014957/13	3578.34	Pick up upto 2350cc		OWNGOODS	7	1390			
AIC/ADD/MTCM/015000/13	9778.4	Pickup over 2350cc		OWN GOODS	16	2494			
AIC/ADD/MTCM/015062/13	1604.52	Pickup over 2350cc	TOYOTA	OWN GOODS	23	1998			
AIC/ADD/MTCM/015320/13	24961.47	General Cartage - Trucks With Trailers	FIAT IVECO	GENERAL CARTAGE	11	13798			
AIC/ADD/MTCM/015497/13	4584.62	Pickup over 2350cc	TOYOTA	OWN GOODS	23	2446			
AIC/ADD/MTCM/015655/13	2426.85	Pickup over 2350cc		OWN GOODS	14	2446			
AIC/ADD/MTCM/015683/13	18646.45	Car Hire - Station Wagon		CAR HIRE	13	4164			
AIC/ADD/MTCM/015784/13	2407.16	Pick up upto 2350cc	TOYOTA	OWN GOODSCARRYI	10	1998			
AIC/ADD/MTCM/015994/13	14700	General Cartage - Isuzu Trucks - NPR		GENERAL CARTAGE	5	4334			
AIC/ADD/MTCM/016111/13	18672.44	General Cartage - Trucks With Trailers	MERCFDES	GENERAL CARTAGE	12	11500			

Table 3. 5 : Claim raw data

Claim Number													
A	B	C	D	E	F	G	H	I	J	K	L	M	
Claim Number	Policy Number	Claim Date	Premium	Total Premium	Claim Amount	Risk Name	Vehicle Make	Vehicle Model	Purpose of the Vehicle	Age of the	Cylinder Capacity	Age of the driver	
2	CLN/AHO/MTCM/004057/14	AIC/ADM/MTCM/030577/13	23-05-2014	79945.39	22280382	3049377	General C EUROTRUCK		GENERAL CARTAGE	1	12880		26
3	CLN/AHO/MTCM/007958/14	AIC/JIM/MTCM/061377/13	08-12-2014	29479.66	8201279	2327300	General Cartage - Trucks		GENERAL CARTAGE	3	12800	no	
4	CLN/AHO/MTCM/005239/14	AIC/ADD/MTCM/012657/13	07-12-2013	71550	22280382	1800000	General C IVECO		GEERAL CARTAGE	7			28
5	CLN/AHO/MTCM/002979/14	AIC/ADD/MTCM/037574/13	20-10-2013	1679401	16156503	1614097	Pickup ovi TOYOTA		OWN GOODS	1	2494	NA	
6	CLN/AHO/MTCM/001040/13	AIC/MRT/MTCM/023544/13	01-10-2013	48000.78	16156503	1577779	Pickup over 2350cc		OWN GOODS	1	2494	no written on notification	
7	CLN/AHO/MTCM/004747/14	AIC/GTR/MTCM/030825/13	20-04-2014	69012.75	16156503	1563923	Pickup over 2350cc		OWN GOODS	0	2494		36
8	CLN/AHO/MTCM/000925/13	AIC/ADM/MTCM/043979/13	16-09-2013	38257.66	22280382	1500000	General C IVECO		GENERAL CARTAGE	2	12880	no	
9	CLN/AHO/MTCM/000849/13	AIC/FFN/MTCM/038602/13	27-08-2013	113023.8	22280382	1448106	General Cartage - Trucks With		GENERAL CARTAGE	3	12880	NA	
10	CLN/AHO/MTCM/001416/13	AIC/BOL/MTCM/016791/13	04-09-2013	154049.9	16156503	1374262	Pickup ovi	0	OWN GOOD CARRYING	1	2494	none	
11	CLN/BOL/MTCM/005132/14	AIC/BOL/MTCM/056575/13	25-07-2014	24065.79	2340054	1330318	General C JAPAN		GENERAL CARTAGE	5	12503	NA	
12	CLN/AHO/MTCM/004385/14	AIC/GFM/MTCM/004232/12	15-06-2014	23050.88	17102584	1308390	General C CHINA		GENERAL CARTAGE	1	9726	NA	
13	CLN/AHO/MTCM/003588/14	AIC/BOL/MTCM/057707/13	03-01-2014	22332.24	17102584	1305000	General Cartage - Tipper Famili		GENERAL CARTAGE	1	12880	NA	
14	CLN/AHO/MTCM/000980/13	AIC/ADD/MTCM/024839/13	23-08-2013	27813.6	17102584	1304450	General Cartage - Tipper Famili		GENERAL CARTAGE	1	9726		32
15	CLN/AHO/MTCM/000713/13	AIC/KOL/MTCM/013981/13	16-08-2013	26013.6	17102584	1272048	General Cartage - Tipper Famili		GENERAL CARTAGE	1	9726	no	
16	CLN/AHO/MTCM/001985/14	AIC/SHA/MTCM/051280/13	01-01-2014	6531.84	527936.6	1152264	Comm Usi GERMAN		AGRICULTURAL USE	7	152	no written on file	
17	CLN/AHO/MTCM/000810/14	AIC/SHA/MTCM/065203/13	16-12-2014	8706.57	527936.6	1112140	Comm Use Combine Harvester		AGRICULTURAL	6	0		37
18	CLN/AHO/MTCM/001857/13	AIC/GFM/MTCM/041908/13	17-12-2013	115169.2	22280382	1101461	General C JAPAN NISSAN		GENERAL CARTAGE	6	12503	not applicable	
19	CLN/AHO/MTCM/000340/13	AIC/NFS/MTCM/024661/13	13-06-2013	35100	2200487	1075817	General Cartage - Truck Over 1t		GENERAL CARTAGE	4	97261		36
20	CLN/AHO/MTCM/006081/14	AIC/BOL/MTCM/014843/13	01-02-2014	261600	2200487	1049528	General C IVECO		GENERAL CATRAGE	7	13798	none	
21	CLN/AHO/MTCM/002020/14	AIC/PIZ/MTCM/015569/13	04-01-2014	23658.34	1262208	1046750	General Cartage - Tipper Mitsi		GENERAL CARTAGE	1	9726		28
22	CLN/ADD/MTCM/004737/14	AIC/ADD/MTCM/043923/13	07-07-2014	13477.15	22280382	1025908	General C IVECO		GENERAL CARTAGE	14	13797	xx	
23	CLN/ADD/MTCM/003973/14	AIC/ADD/MTCM/048381/13	20-05-2014	43200	3651379	1011325	Own Goods Trucks		OWN GOODS	8	11051		28
24	CLN/AHO/MTCM/001884/13	AIC/GTR/MTCM/015190/13	21-12-2013	57500	16156503	1004728	Pickup over 2350cc		OWN GOODS CARRYING	5	2494		28

Note: during data collection from Awash Insurance database (underwriting and claim data) all the fields are not included. Only selected fields or attributes are selected.

3.5.2 Data Pre-Processing

Data pre-processing is preparing the dataset for further analysis which is suitable for machine learning tasks to be performed. In the raw data usually there are many irregularities in the dataset that might distort the output of the experiment. The pre-processing phase consists of data cleaning, dataset splitting, attribute selection, data integration and data formatting.

3.5.2.1 Data Cleaning

Data with outliers or missing values have been removed from the dataset so that the output will give representative information for decision making. As the data in Awash Insurance Company database contain many missing values, inconsistent data, there are many records which have been corrected in consultation with underwriters and branch managers and few of them are removed from the record so as to make it more convenient for machine learning. Finally 52,831 records have been taken as complete and consistent records for further experimentation. The datasets were checked for completeness and correctness of the required attributes using excel file filtering and sequential arranging programs before analysis and prediction for the quality assurance of the experiment under process.

3.5.2.2 Dataset Splitting into Training and Testing data

As 80/20 data splitting for training and testing data is a common practice of splitting ratio for samples of a moderate size in the machine learning applications(Alakwaa, Chaudhary and Garmire,2017), this research paper also follows the 80% training data and 20% testing data partitioning technique.

3.5.3 Data integration

The data collected from two data tables namely underwriting and claim data tables are exported to excel format and the column arrangement is done in separate files. Then every claim incurred in the claim data table moved to the corresponding policy number in the underwriting data table. The fact behind is every policy number in the claim data table is available in underwriting data table but the reverse is not true.

3.5.4 Data formatting

The data is converted to excel format. The data clearing starts from Risk Item name. There are well known risk item names as detailed in Table 4.1 and all the vehicles listed according to their category. For each risk item name again categorization based on Vehicle- make has been done. This is usually identifying the country in which the vehicles are made in. taking the vehicle make as a baseline; another categorization is done based on the purpose of the vehicle. This is to identify the purpose of the vehicle used for. The same group of vehicles again categorized base on the Age of the vehicle. Finally, they are sub categorized again based on the Carrying capacity of the vehicles. In summary each Risk Item sub-Categorized is presented as follows:

Risk Item Name-> Vehicle Make-> Purpose of the vehicle->Age of the vehicle->Carrying Capacity of the vehicle. And therefore, each Risk Item Name has passed all these sub categorization. Except Risk Item Name, there were many wrong data entry as they are user-entry data. To complete and correct the data intensive contact with branch user where these policy have been issued has been done especially for recent policies. Those old policies with no relevant evidence to correct the data have been removed from the dataset. To facilitate this data manipulation excels filtering and ordering system has been used to visualized and take appropriate measure easily.

3.5.5 Maintaining Balances

The data set under experiment is to be classified into a category of acceptable claim ratio which is less than 66%, is applied standard claim ration that is fixed by company executive management with consent of board of directors. This standard rate is similar throughout the company's branches or underwriting units. This rate is usually known as standard rate. For the purpose of this study it simply represented as "Standard". The other category is records with high claim ratio or More than 66% claim ratio. In the company risks with high claim ratio is subjected to premium loading. For the purpose of this experiment, simply it is termed as "Loading". Out of the 52,831 dataset 29,418 records categorized under "Standard" and 23,413 records categorized under "Loading" category.

3.5.6 Attribute Selection

An attributes selection method was based on the correlation between the data features or attributes of the chosen dataset (Kamel and Nour, 2018).

During underwriting there are many data to be captured. But all these data are not important from the purpose of this paper perspective. As it is stated earlier, the purpose of this paper is to identify Risk items with high claim cost or loss making and profit making. Most of the attributes captured during underwriting do not serve this purpose. For instance the following attributes are not included in the experiment. Proposal number, Plate Number, Chassis Number, engine number, number of vehicle covered, Duty free, policy type, Rate type, Business source, Period of insurance, customer name, Customer Information file, Host Branch etc. On the other hand the following attributes are considered important for the purpose of this paper. Policy name(represented by PC), premium collected, claim paid, claim history, Risk Item , Vehicle make, purpose of the vehicle, Age of the vehicle carrying capacity of the vehicle, and Premium (premium rate to be applied) and included into the experiment. From motor insurance point of view, age and gender of the driver are very relevant parameter which should be taken into consideration. Unfortunately these parameters are not available in underwriting data table as the driver of the vehicle can be anyone other than the owner of the vehicle.

Taking into account the dataset under experiment, the names and values of the attributes have been changed into some generic symbols for the sake of simplicity to have a more accurate representation of the variables.

Table 3. 6 : Attributes selected for experiment.

Attribute	Description	Data Type
Policy Code	It represents the policy number since the format of policy number is not convenient for model building it is represented by policy code which is basically two values; either motor commercial or motor private.	Character
Premium Amount	The amount of premium received by Insurance company in exchange of insurance coverage promised.	Numeric
Claim Amount	The amount of Claim paid for specific policy	Numeric
Claim History	It is a value which indicates whether claim happened or not to specific policy of the same value in each attributes.	string
Risk Item	It is a name of each risk Item in the dataset	string
Vehicle Make	Vehicle make is the country where the vehicle is manufactured.	string
Purpose of the Vehicle	It is to indicate for what purpose the vehicle is used for which insurance coverage is valid.	string
Age of the vehicle	This is age of the vehicle since manufactured.	Numeric
Carrying capacity	It indicates the allowable carrying capacity of the vehicle set by the manufacturer.	Numeric
Premium (how to calculate premium)	This is the dependent variable that this experiment is to be conducted to know how to charge the premium for specific risk item (either using “Standard rate” or “Loading”.	string

Note: “Premium Amount” described above is the actual premium amount that Awash insurance company has collected from the insured. But the dependent variable termed as “Premium” represents a proposed premium to be collected based on the suggested rate type during underwriting (purpose of this research paper).

3.6 Data Transformation

The actual data in the dataset are not suitably for prediction techniques as it is. Therefore, it needs some transformation which makes the pattern of the data easy for the machine learning process in order to make sensible prediction. The dataset in this experiment and analysis contains categorical values that are transformed to binary values or factors.

The values of an attributes are changed as follows.

Table 3. 7 : Attributes Transformation

Attributes change	
Policy Code	PC
Premium Amount	PA
Claim Amount	CA
Claim History	CH
Risk Item	RI
Vehicle Make	VM
Purpose of the Vehicle	PoV
Age of the vehicle	AoV
Carrying capacity	CC
Premium(premium rate to be user)	PRM

“Policy Code”, each policy Number under this study is either motor commercial or motor private with branch code and year of insurance. They are represented as follows.

Table 3. 8 : Policy Code Transformation

Policy code		
Risk class	transformed	Binary value
Motor commercial	MTCM	0
Motor Private	MTPV	1

“Claim History”, policies with claim history is represented with “Yes” and policies without claim represented with “No”

Table 3. 9 : Claim History Transformation

Claim History (CH)		
Claim History	Transformed	Binary Value
Policy without claim	No	0
Policy with claim	Yes	1

“Risk Item”, currently, Awash insurance gives insurance coverage for many risk items. For the sake of this study the data is transformed as follows. Factor is a variable of categorical value which contains discrete values that is important to facilitate pattern recognition in data mining process.

“Vehicle Make”, different vehicles are produced in difference countries. The specification and quality of the vehicles also varies from country to country. The data transformation of the vehicle make is done as follows.

Table 3. 10 : List of vehicle makes

Vehicle Make (VM)		
Country	Transformed To	Factor value
BRAZIL	CTR1	1
CHINA	CTR2	2
ENGLAND	CTR3	3
FRANCE	CTR4	4
GERMANY	CTR5	5
INDIA	CTR6	6
ITALY	CTR7	7
JAPAN	CTR8	8
KOREA	CTR9	9
NETHERLANDS	CTR10	10
SPAIN	CTR11	11
USA	CTR12	12

“Purpose of the vehicle”, purpose of the vehicle is for what the customer is using the vehicle and it is transformed as follow.

Table 3. 11 : List of purposes the vehicles are user for.

Pupose of the Vehicle(PoV)		
Purpose of the Vehicle	Transformed to	Factor value
AGRICULTURAL USE	PRP1	1
AMBULANCE	PRP2	2
CAR HIRE	PRP3	3
FUEL TANKER	PRP4	4
GENERAL CARTAGE	PRP5	5
GENERAL CARTAGE TIPPER	PRP6	6
LEARNER	PRP7	7
METER TAXI	PRP8	8
OWN GOODS	PRP9	9
OWN SERVICES	PRP10	10
PRIVATE USE	PRP11	11
PUBLIC TRANSPORT	PRP12	12
SPECIAL PURPOSE	PRP13	13
TAXI	PRP14	14
TOUR AND TRAVEL	PRP15	15

“Age of the Vehicle”, the age of the vehicle is categorized into different groups. As a policy for new vehicles the company gives model discount and for old vehicles there is loading that means subjected to additional premium. It is categorized as follows.

Table 3. 12 : Group of Age of the vehicles

Age of the Vehicle(AoV)		
Age Group	Transformed to	Factor Value
0-3	MD1	1
4-10	MD2	2
11-15	MD3	3
16-20	MD4	4
21-25	MD5	5
>26	MD6	6

“Carrying Capacity” carrying capacity of a vehicle is the standard amount allowable weight to be transported by the vehicle. The premium charging depends on the carrying capacity of the vehicles. And therefore, it is categorized as follows.

Table 3. 13 : Carrying Capacity of the Vehicles

Carrying Capacity(CC)		
Carrying Capacity	Transformed to	Factor Value
<1000	CCP1	1
1001-1200	CCP2	2
1201-1400	CCP3	3
1401-1650	CCP4	4
1651-2000	CCP5	5
2001-2350	CCP6	6
2351-3050	CCP7	7
3051-3650	CCP8	8
3651-4400	CCP9	9
>4400	CCP10	10

“Premium”, as per the discussion made with the underwriting and branch operation manager, Premium is an amount to be collected from a customer to the insurer in exchange of insurance coverage provided by the insurance company. It is a dependent variable on the claim ratio of the policy. It can be either “Standard” if the claim ratio is within acceptable range or “Loading” if the claim ratio is beyond the acceptable range

Table 3. 14 : Premium to be calculated based on the risk ration

Premium to be collected(RPM)		
Risk Ratio	Transformed to	Binary Value
< 66%	Standard	0
>=66	Loading	1

The dataset under experiment looks like as follow in RStudio

Table 3. 15 : Dataset in R Studio

	PN	PC	PA	CA	CR	CH	RN	VM	PoV	AoV	CC	PRM
1	AIC/GRJ/MTPV/307377/18	1	4400	0	0%	0	1	8	2	1	7	0
2	AIC/GTR/MTPV/016692/13	1	8800	0	0%	0	1	8	2	1	7	0
3	AIC/PIZ/MTPV/242860/17	1	6600	0	0%	0	1	8	2	1	7	0
4	AIC/SAB/MTPV/271876/17	1	5205	0	0%	0	1	8	2	1	7	0
5	AIC/KZC/MTPV/198101/16	1	8885	5000	89%	1	1	8	2	2	7	1
6	AIC/KZC/MTPV/191774/16	1	27082	6900	79%	1	1	8	2	1	9	1
7	AIC/ADD/MTPV/309060/18	1	13703	7000	84%	1	1	8	2	1	9	1

Showing 1 to 8 of 52,831 entries, 12 total columns

3.7 Framework for Guiding Data Mining Tasks

This research paper followed Cross-Industry-Standard-Process for Data Mining (CRISP-DM) process phases to build a predictive model that can help the insurance industry in predicting or classifying risk items according to the volume of risk ratio they have during underwriting process in Awash Insurance Company either as “Standard” risk items with acceptable risk claim ratio and “Loading” which are risks with high claim ratio.

3.7.1 Business understanding

This initial phase focuses on understanding the research objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives (Sastry and Babu, 2013).

Awash insurance company is one of the first few pioneer private insurance companies in Ethiopia launched following the liberalization of the financial sector in 1994. For the last 6 years it is the leading insurance company in the industry from private insurance companies except 2017/18 fiscal year. Currently it has 48 branches and 6 contact offices in the country.

From the interview and discussion made with Branch managers, Claim managers as well as underwriting and claim officers, the motor class of business takes the lion share of the insurance business in the industry in general and awash insurance company in particular. According to the claim manager motor business is not attractive as most of the claim payments go to motor class of business from year to year which is justified by the claim report of the company. For example from July1, 2017-June30, 2018 the claim ration of Awash Insurance Company is 65.2% for Motor Private Own Damage Policy and 74.6% for Motor Commercial (AIC CONSOLIDATED CLAIMS ANALYSIS, 2018). But no insurance company can avoid it since it is the major line of business through which they can get other class of business like fire and burglary, marine, engineering etc. which are attractive or major profit making class of businesses.

On the other hand from the discussion with branch managers, it is not possible to say motor class of business is a loss making business; because branch managers confirmed that there a lot NCD(No Claim Discount) cases to be provided to their customers every year. That means some risk items are less claim cost prone and others are highly claim cost prone from motor class of business. Currently, there is no way to differentiate risk items either as attractive or non-attractive groups. It is too difficult to do this manually as the data is voluminous. Moreover, it is not easy to associate different parameters to a given risk item to say this risk item with theses parameters is high or low risk ration. In general, since there are huge amount of motor risk items in day to day insurance business transaction in underwriting and claim process, it is very difficult to identify which risk items are particularly causing higher amount to claim cost manually. Therefore, the claim report produced by claim department is as traditional as to say from the total claim paid; the big ratio goes to motor class of business which is very generic.

3.7.2 Data Understanding

The data understanding phase starts with initial data collection and proceeds with activities that enable us to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information (Shafique and Qaiser, 2014).

In this research paper the data is collected from Awash insurance company database which includes underwriting and claim data. Integrating these data from underwriting and claim tables enables the researcher get claim ration which is one of the dependent attributes of training and testing data. Accordingly for this study the data is collected from all the branches of Awash Insurance Company containing 52, 831 records with 10 attributes out of which one of the attribute is dependent fields represented as either acceptable claim ration (“Standard”) or high claim ratio (“Loading”).

3.7.3 Data Preparation

This sub-topic deals with preparing the data that an experiment is intended to be performed on.

Data preparation is a fundamental stage of data analysis. The reason for its fundamental base is, there is a possibility of low quality of information to be available in the data source which requires transformation to high quality data for further analysis (Alasadi and Bhaya, 2017).

This is dedicated to detailed data preparation and discussed each and every activity of data preparation phase which covers all activities needed to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Finally, the dataset is saved as .csv format. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools (Sastry and Babu, 2013). Therefore all the data preparation processes are applied on the data collected from Awash insurance company's database or from underwriting table and claim table starting from 2012 up to 2019 for the purpose of intended experiment. From the records retrieved from the databases, there are many records without required field values and unreliable values which should be avoided from the experiment. Therefore, the data used under this research paper considers only complete in terms of required fields, values and configured rate type which is common rate type for premium calculation throughout the company.

3.7.4 Modeling

This is the fourth phase of CRISP-DM process selection and application of various modeling techniques. In this phase, various modeling techniques are selected and applied, and their parameters are set to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.

Using the above data taken from Awash Insurance Company database, this paper uses some machine learning algorithm to learn certain patterns from the data provided in training data to predict the likely category of each risk item along with all the selected testing parameters when the testing data is applied on. As SVM generally are capable of delivering higher performance for small and medium datasets (in our case 52,831 records) in terms of classification accuracy

(Srivastava and Bhambhu, 2010), it is chosen for this experiment. Moreover, it is memory efficient. The output we expect from this experiment is two which is “Standard” and “Loading” and therefore, as Logistic regression model is suitable to predict dichotomous outcomes (Jihye Jeon, 2015); it is also chosen as good algorithm for our dataset. On the other hand as Naives Bayes algorithm is relatively simple to understand, can be trained easily and faster to predict classes (Kaviani and Dhotre, 2017), it is employed in this study too. Moreover, Rstudio is employed as it is widely used among data miners for developing statistical software and data analysis (Tejashree and Sawant, 2016). Some of the fields in the dataset of this study contains binary data and other are transformed to numeric data type for which Rstudio is very suitable as well. According to Tejashree and Sawant (2016), R’s popularity substantially increased in recent years as a result of its ability to provide quality with ease of statistical and graphical techniques.

3.8 Data Visualization

Data visualization is also known as information visualization which makes messages or information that cannot be touched, smelled or tasted last in time (Sadiku ,Shadare, Musa and Akujuobi,2016).For this study, visual analysis conducted on the dataset under experiment, R programming is employed to build visualization of the attributes.

One of the variables under this experiment is policy type. As it is stated under this study only two policy types is included. These are private vehicles which are usually represented as “MTPV” and Commercial Vehicles which are represented as “MTCM”. As it is depicted in Figure 3.6 below motor private vehicles outnumber motor commercial vehicles.

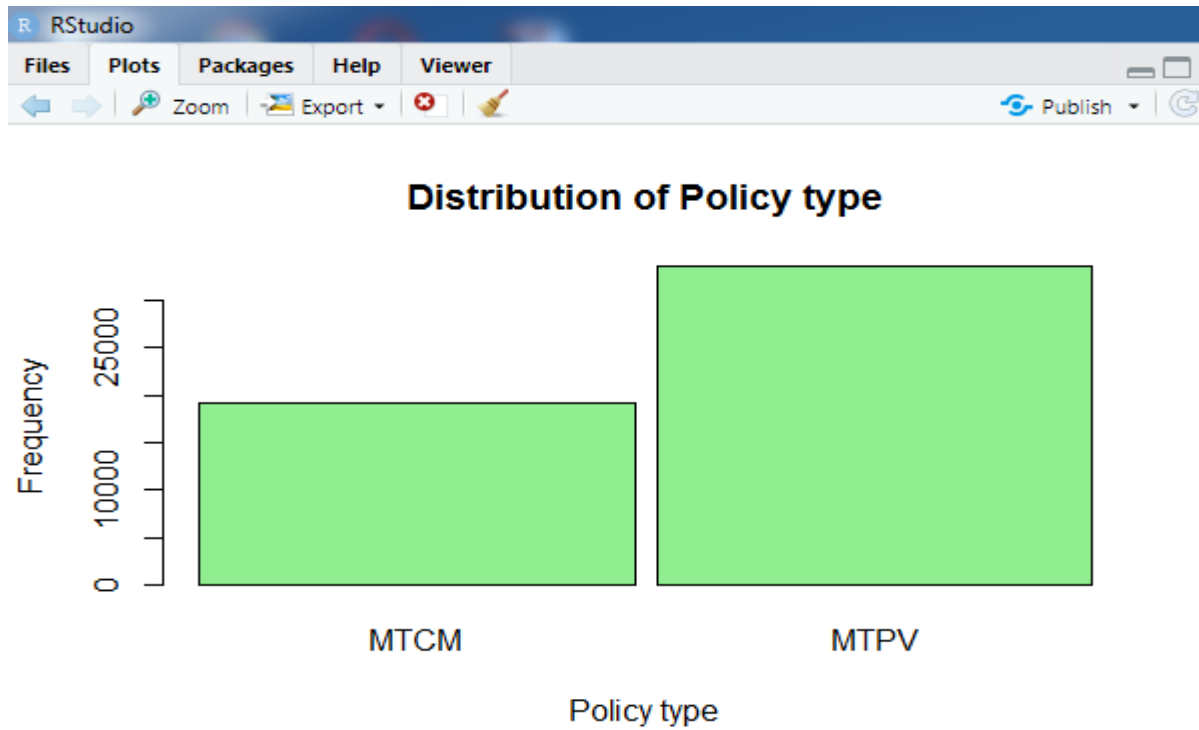


Figure 3. 6: Policy type distribution

The other variable taken as important variable is premium collected which is independent of any variable in the experiment. As per discussion made with senior underwriters in Awash insurance the premium collected cannot be less than Br. 500 which is minimum and rarely around Br. 100,000. This is depicted in

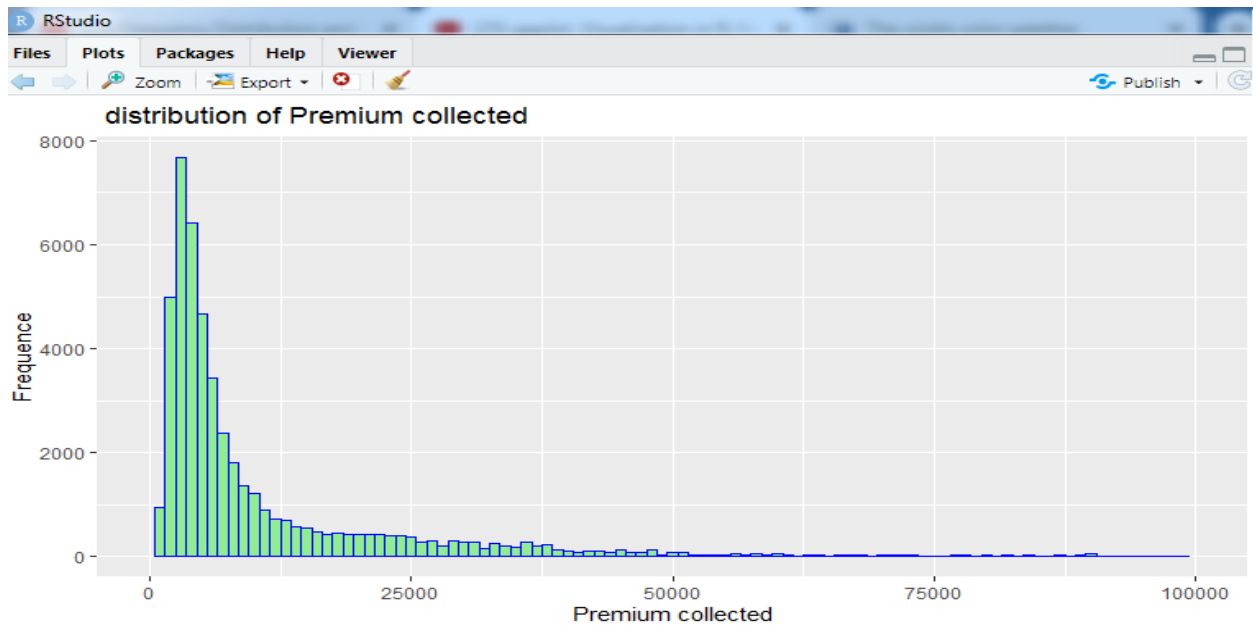


Figure 3. 7 : Premium collected distribution.

Note: the Bar chart depicted at figure 3.7 above show majority of the premium collected lies between Br 5, 000 which is the minimum premium set for private vehicles and 10,000 as one can understand from the graph. The reason is most of the vehicles in the country are old model and less priced which is directly related to the premium to be collected.

Claim is the amount that the insurance company pays when unintentional accident happens to a property under insurance coverage. Since it is not easy to include all claims paid as the number is too big, this experiment includes claims paid up to 100, 000 per vehicle which is depicted under

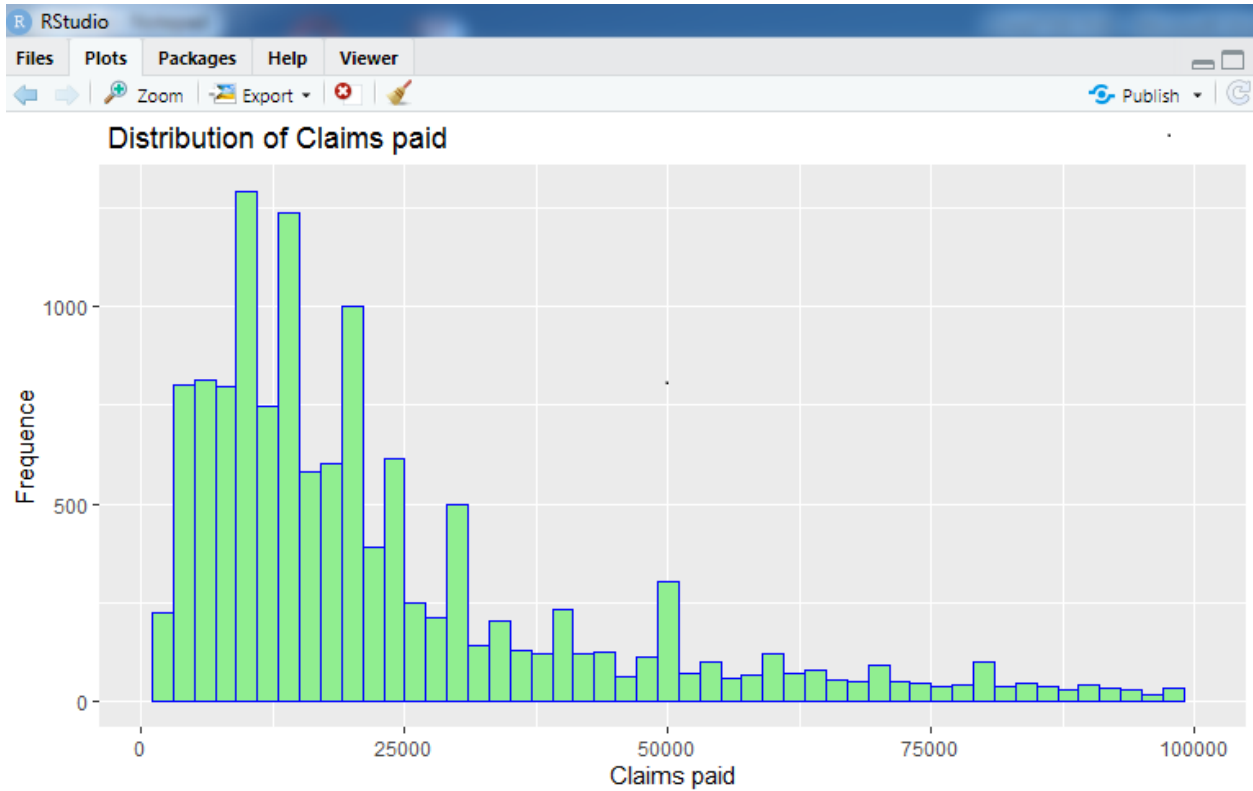


Figure 3. 8: Claim paid distribution

Note: Figure 3.8 above shows what claim amount are paid so far frequently. Since premium collected has direct relation to the claim amount to be paid, the claim density depicted above also shows high claim cost which is highly frequent between Br. 5,000 and 10, 000.

The other one is claim history which is a mechanism of checking whether claim has been made to a particular policy before or not. As it is clear from Figure 3.9 below, the frequency or the number of claim made is less than the number of no-claims. The claim amount to be paid is limited to the sum insured (current Market value of each vehicle) which is the maximum claim amount for each vehicle in case of total damage

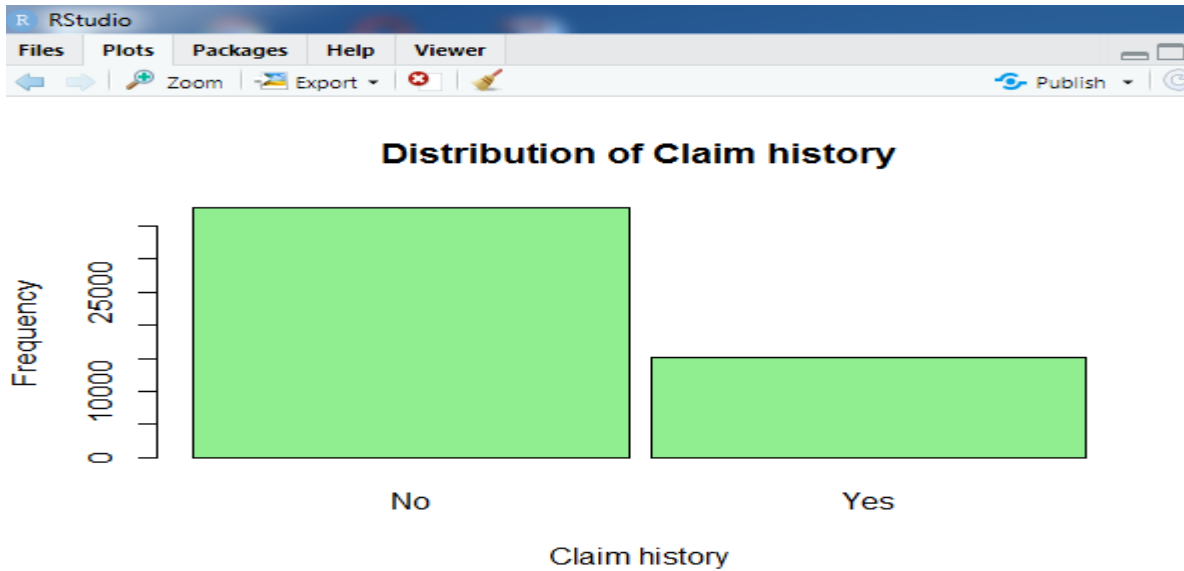


Figure 3. 9: Claim history distribution

Risk item distribution is difficult to display graphically unless some kind of representation is done, because some risk item's name is too long to display on the graph. Therefore, it is represented as follows.

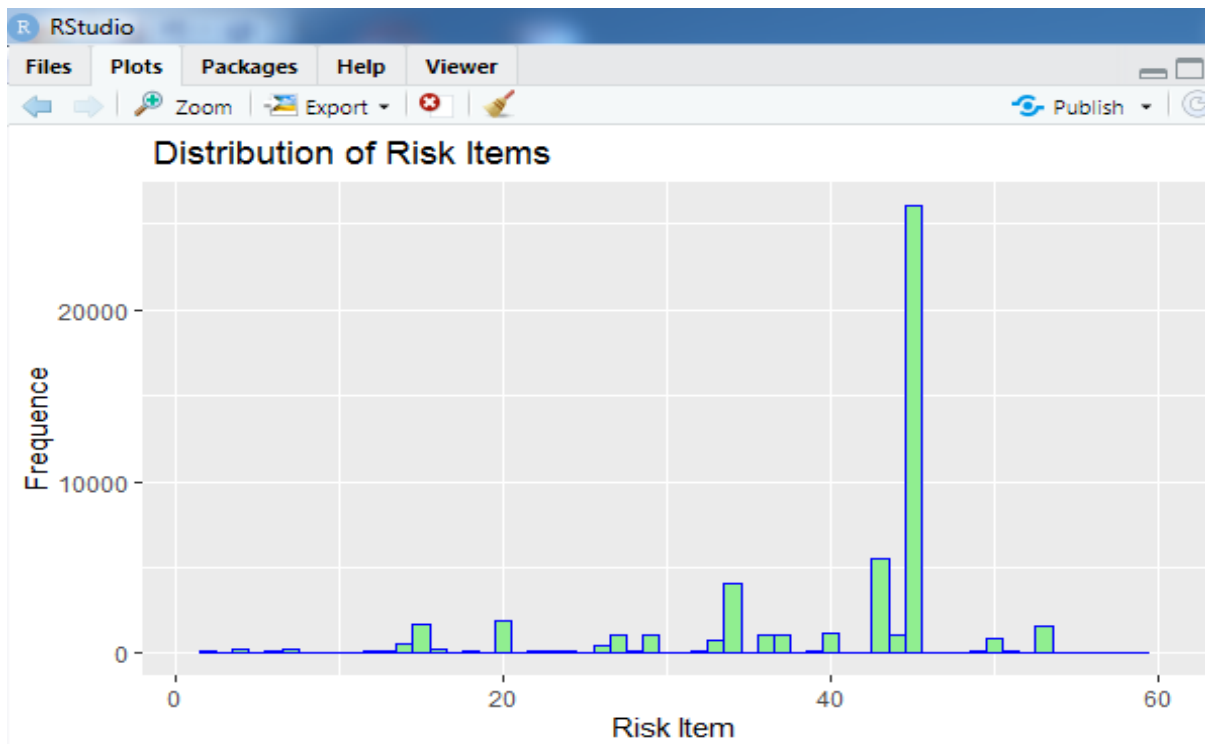


Figure 3. 10 : Risk Items distribution with number representation

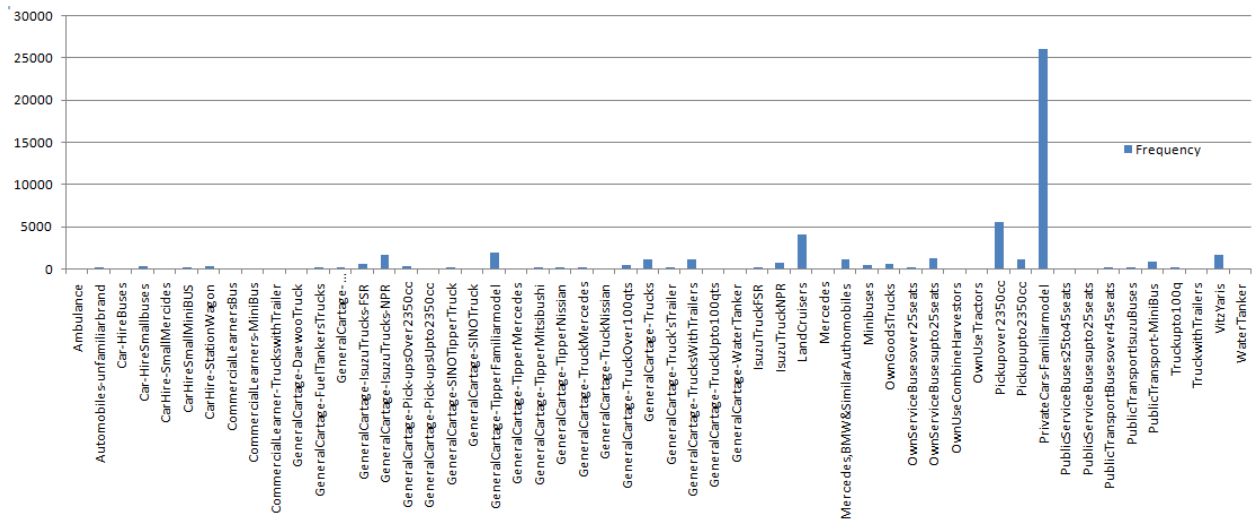


Figure 3. 11 : Risk Items distribution with actual risk items name

Note: As the number of Risk Item Names is too large to represent in RStudio it is attempted to represent with numerical values. The highest frequent Risk Item Name got cover in Awash Insurance Company is between 40 and 60. But it is difficult to spot particularly which Risk Item name is it. Fig. 4.6 solve this problem. It shows exactly the Risk Item name which is “Private cars familiar model” which is followed by pickup over 2350cc. Basically, fig 4.5 and fig.4.6 are used for the same purpose in different format.

The country in which the vehicles manufactured, which has been provided with insurance coverage in Awash insurance is known as vehicle-make. It is clearly show that Japan made vehicles dominate next to Chinese as depicted in fig 4.7 below.

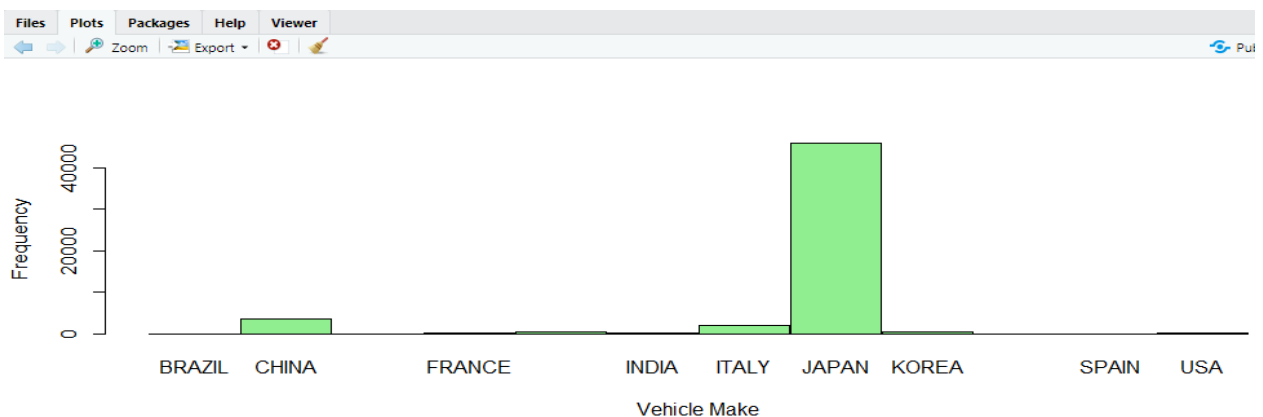


Figure 3. 12 : Vehicle- make graphical representation

The independent variable used under this experiment is Purpose of the vehicle. For making graphical visualization easily understandable, numerical values are used to represent the purpose of the vehicles as follows. As it is depicted in fig4.8, the Private use vehicles outweigh by large the others

Table 3. 16 : Numerical representation of purpose of the vehicles.

Purpose of the Vehicle	Represented by
AGRICULTURAL USE	1
AMBULANCE	2
CAR HIRE	3
FUEL TANKER	4
GENERAL CARTAGE	5
GENERAL CARTAGE TIPPER	6
LEARNER	7
METER TAXI	8
OWN GOODS	9
OWN SERVICES	10
PRIVATE USE	11
PUBLIC TRANSPORT	12
SPECIAL PURPOSE	13
TAXI	14
TOUR AND TRAVEL	15

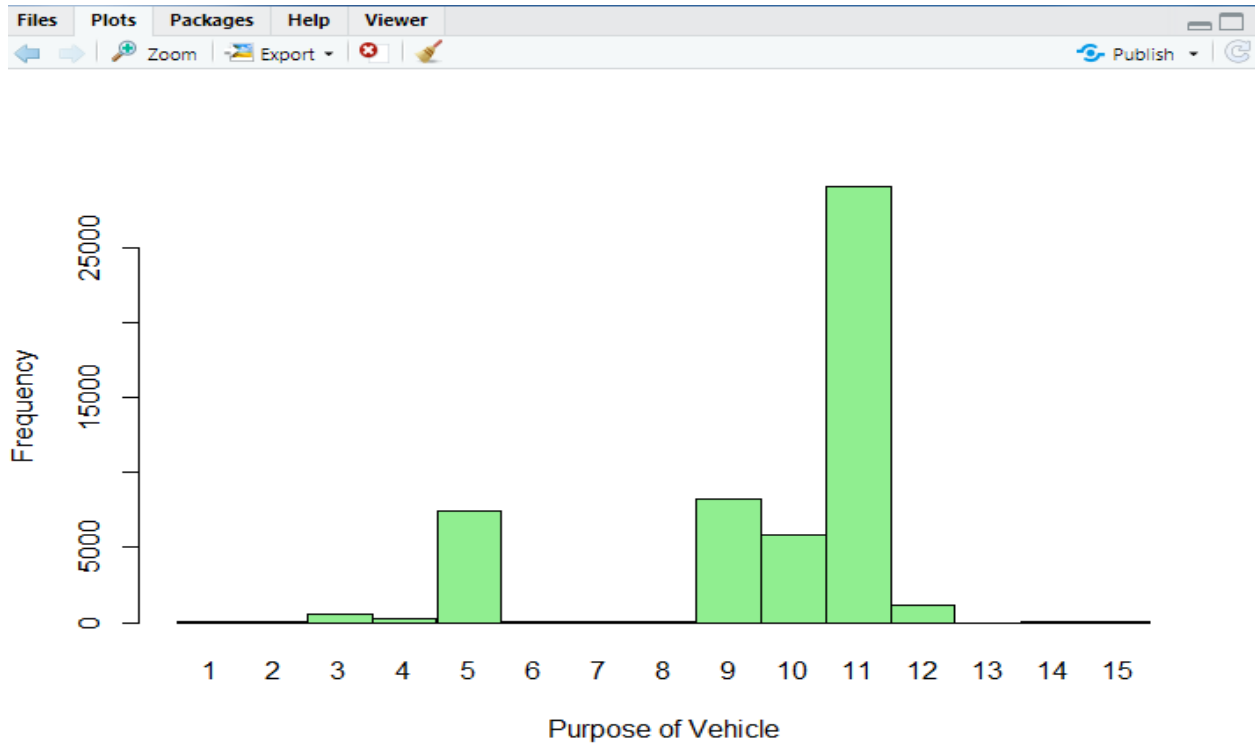


Figure 3. 13 : Number of vehicles by purpose

Age of the vehicle is the other independent variable which is categorized into groups as it is a rating parameter. The categorization is used for either model discount or loading based on the age of the vehicle’s category. In Awash insurance company the grouping is done as follows. For ease use of graphical display, the following representation is done in table 4.3 below. Age of the Vehicle

Table 3. 17 : Age group of the vehicles under experiment

Age group	Represented by
0-3	1
4-10	2
11-15	3
16-20	4
E 21-25	5
>26	6

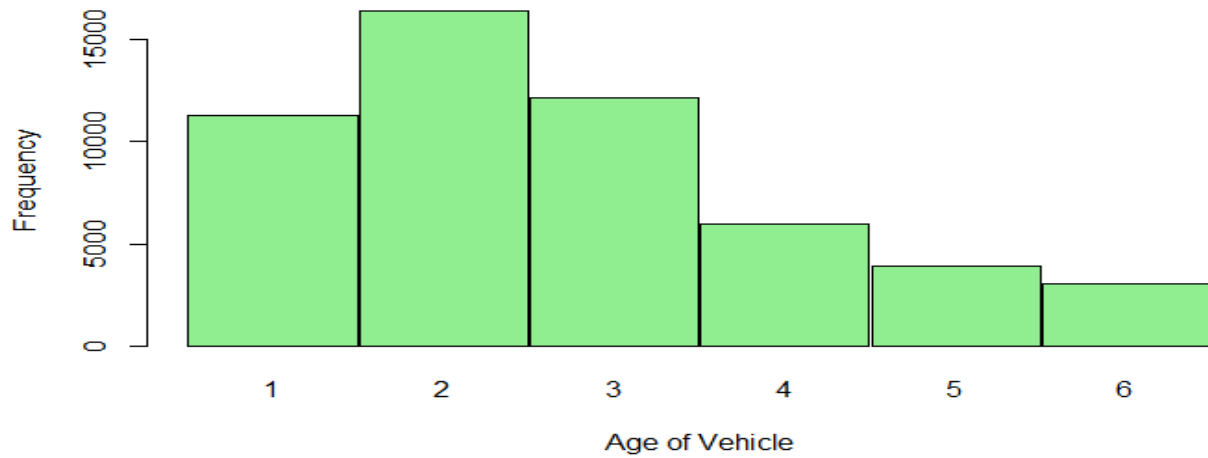
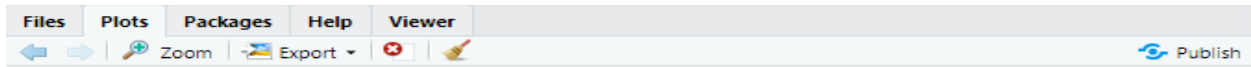


Figure 3. 14 : Number of vehicles by age of the vehicles under the experiment

Carrying Capacity of the vehicle is also the rating parameter which is an independent variable. In Awash insurance company the carrying capacity of each vehicle is categorized as depicted in Table 3. 18 below.

Table 3. 18 : Numerical representation of CC.

Carrying Capacity	
CC	Represented by
<1000	1
1001-1200	2
1201-1400	3
1401-1650	4
1651-2000	5
2001-2350	6
2351-3050	7
3051-3650	8
3651-4400	9
>4400	

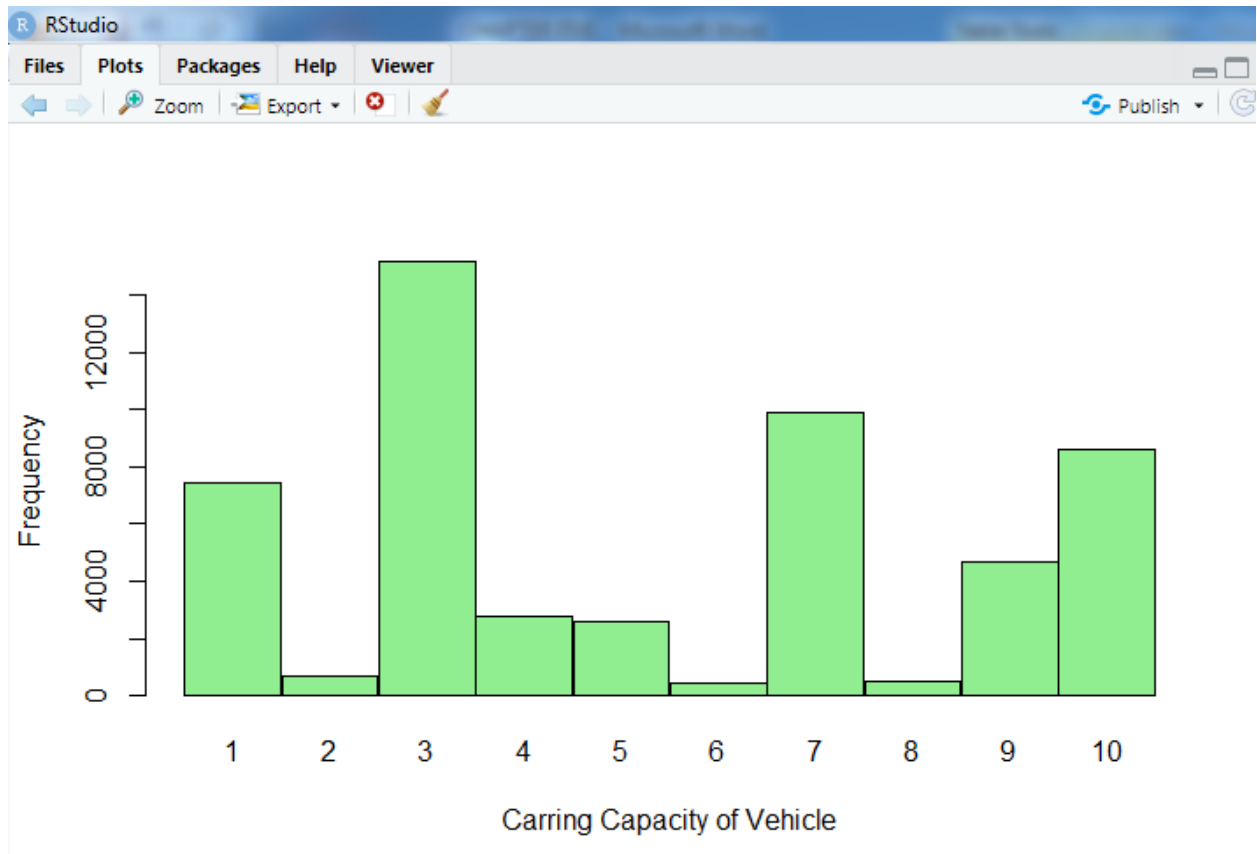


Figure 3. 15 : Number of Vehicles by Carrying Capacity of the vehicles under experiment

CHAPTER FOUR

4 EXPERIMENT AND DISCUSSION OF THE OUTPUT

4.1 Introduction

Under this chapter, different experiments will be conducted on the dataset Prepared as described in chapter three under 3.4.3 (Data Pre Processing). Before data cleaning, it is obvious that the number of records exceed the number of records in the dataset after data cleaning as described under 3.4.4. After data cleaning, the total number of record in the dataset is 52,831. Among the total dataset 80% if of it is used to train the algorithm of the model and 20% of it has been used for testing the performance of the model selected for the experiment.

Accordingly the experimentation is conducted using machine learning models which are supervised vector Machine (SVM), Logistic Regression and Navies Bayes models as discussed in chapter three under modeling section (3.3.4). Each model will be evaluated and the best performing model will be chosen as best machine learning model for the objective of the experiment under discussion. R Studio is employed for the experiment to be carried out. Under this section several activities will be conducted regarding running and evaluating model building experiments, selecting the best and appropriate model, and providing explanations on the selected model are main activity of this chapter.

4.2 Experiment

4.2.1 Support Vector Machine (SVM) modeling

Under this experiment prediction model was built using Support Vector Machine (SVM) algorithm. A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. (Noel Bambrick, 2019). As it is stated under 3.4.5.2 (Data formatting) this study intended to assist underwriters during underwriting process to determine how to calculate the premium for risk items requesting for insurance coverage. The values are binary which is either loading represented by '1' or Appling standard rate simply "Standard" which is represented by '0'. Out of the 52,831 dataset records 29,418 records categorized under "Standard" category and 23,413 records categorized under "Loading"

category which of the total dataset. Since the two classes are fairly balanced, the actual dataset is preferred to be taken to the experiment as it is. Importing dataset from csv format to RStudio

```
>SVMaic<-read.csv(file.choose(),sep = ",")
```

	PN	PC	PA	CA	CH	RN	VM	PoV	AoV	CC	PRM
1	AIC/GRJ/MTPV/307377/18	1	4400	0	0	1	8	2	1	7	0
2	AIC/GTR/MTPV/016692/13	1	8800	0	0	1	8	2	1	7	0
3	AIC/PIZ/MTPV/242860/17	1	6600	0	0	1	8	2	1	7	0
4	AIC/SAB/MTPV/271876/17	1	5205	0	0	1	8	2	1	7	0
5	AIC/KZC/MTPV/198101/16	1	8885	5000	1	1	8	2	2	7	1
6	AIC/KZC/MTPV/191774/16	1	27082	6900	1	1	8	2	1	9	1
7	AIC/ADD/MTPV/309060/18	1	13703	7000	1	1	8	2	1	9	1
8	AIC/BOL/MTPV/244734/17	1	8269	8000	1	1	8	2	2	7	1
9	AIC/BOL/MTPV/325165/18	1	10000	9200	1	1	2	2	1	6	1
10	AIC/FFN/MTPV/301691/18	1	10516	10000	1	1	8	2	1	9	1
11	AIC/GTR/MTPV/136671/15	1	13945	10350	1	1	7	2	1	6	1
12	AIC/ADD/MTPV/254939/17	1	13561	11632	1	1	8	2	1	9	1
13	AIC/GTR/MTPV/136676/15	1	12720	12000	1	1	7	2	2	9	1
14	AIC/KZC/MTPV/269237/17	1	10867	12383	1	1	8	2	1	9	1

Figure 4. 1 : Dataset for SVM model snap shot

As it is depicted in figure 4.1 above the dataset consists 52,831 records with 11 parameters. But since Policy Name (PN) is too long string which consumes much memory in R processing, it is categorized as motor commercial (MTCM) or motor private (MTPV) and represented by policy code (PC). Therefore, in this experiment 52, 831 records and 10 parameters are used.

```
R RStudio
> str(SVMaic)
'data.frame': 52831 obs. of 11 variables:
 $ PN : Factor w/ 21379 levels "AIC/22M/MTCM/027913/1
15229 2042 6151 6388 10281 ...
 $ PC : int 1 1 1 1 1 1 1 1 1 1 ...
 $ PA : Factor w/ 14907 levels "1000","10000",...: 973
13737 2 353 ...
 $ CA : int 0 0 0 0 5000 6900 7000 8000 9200 10000 .
 $ CH : int 0 0 0 0 1 1 1 1 1 1 ...
 $ RN : int 1 1 1 1 1 1 1 1 1 1 ...
 $ VM : int 8 8 8 8 8 8 8 8 2 8 ...
 $ PoV: int 2 2 2 2 2 2 2 2 2 2 ...
 $ AoV: int 1 1 1 1 2 1 1 2 1 1 ...
 $ CC : int 7 7 7 7 7 9 9 7 6 9 ...
 $ PRM: int 0 0 0 0 1 1 1 1 1 1 ...
```

```

RStudio
> SVMa1c$PA<-as.numeric(SVMa1c$PA)
> SVMa1c$CA<-as.numeric(SVMa1c$CA)
> SVMa1c$PC<-as.numeric(SVMa1c$PC)
> SVMa1c$CH<-as.numeric(SVMa1c$CH)
> SVMa1c$RN<-as.numeric(SVMa1c$RN)
> SVMa1c$VM<-as.numeric(SVMa1c$VM)
> SVMa1c$PoV<-as.numeric(SVMa1c$PoV)
> SVMa1c$AoV<-as.numeric(SVMa1c$AoV)
> SVMa1c$CC<-as.numeric(SVMa1c$CC)
> SVMa1c$PRM<-as.factor(SVMa1c$PRM)
> str(SVMa1c)
'data.frame': 52831 obs. of 11 variables:
 $ PN : Factor w/ 21379 levels "AIC/22M/MTCM/027913/13",...: 12940 13195 18597 19105 152
15229 2042 6151 6388 10281 ...
 $ PC : num 1 1 1 1 1 1 1 1 1 1 ...
 $ PA : num 9738 14118 12272 10791 14173 ...
 $ CA : num 0 0 0 0 5000 6900 7000 8000 9200 10000 ...
 $ CH : num 0 0 0 0 1 1 1 1 1 1 ...
 $ RN : num 1 1 1 1 1 1 1 1 1 1 ...
 $ VM : num 8 8 8 8 8 8 8 8 2 8 ...
 $ PoV: num 2 2 2 2 2 2 2 2 2 2 ...
 $ AoV: num 1 1 1 1 2 1 1 2 1 1 ...
 $ CC : num 7 7 7 7 7 9 9 7 6 9 ...
 $ PRM: Factor w/ 2 levels "0","1": 1 1 1 1 2 2 2 2 2 2 ...

```

Figure 4. 2 : Dataset structures for SVM model snap shot

Note: For better and meaningful prediction the values of attributes are converted to an appropriate data type as follows.

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> summary(SVMa1c)
      PN          PC          PA          CA
AIC/FFN/MTPV/077879/14: 244  Min.   :0.0000  Min.   :    1  Min.   :    0
AIC/BOL/MTPV/033879/13: 220 1st Qu.:0.0000 1st Qu.: 5029 1st Qu.:    0
AIC/BOL/MTPV/048887/13: 203  Median :1.0000  Median : 8015  Median :    0
AIC/BOL/MTPV/070967/14: 163  Mean   :0.6362  Mean   : 7819  Mean   : 21491
AIC/BOL/MTCM/173727/15: 101 3rd Qu.:1.0000 3rd Qu.:10766 3rd Qu.:  7132
AIC/BOL/MTCM/122062/14:  94  Max.   :1.0000  Max.   :14907  Max.   :20213000
(Other)                :51806
      CH          RN          VM          PoV          AoV
Min.   :0.0000  Min.   : 1.00  Min.   : 1.000  Min.   : 1.000  Min.   :1.000
1st Qu.:0.0000 1st Qu.:34.00 1st Qu.: 8.000 1st Qu.: 9.000 1st Qu.:2.000
Median :0.0000  Median :45.00  Median : 8.000  Median :11.000  Median :2.000
Mean   :0.2855  Mean   :39.33  Mean   : 7.507  Mean   : 9.607  Mean   :2.699
3rd Qu.:1.0000 3rd Qu.:45.00 3rd Qu.: 8.000 3rd Qu.:11.000 3rd Qu.:3.000
Max.   :1.0000  Max.   :54.00  Max.   :12.000  Max.   :15.000  Max.   :6.000

      CC          PRM
Min.   : 1.00  0:29418
1st Qu.: 3.00  1:23413
Median : 5.00
Mean   : 5.35
3rd Qu.: 9.00
Max.   :10.00
> |

```

Figure 4. 3 : Dataset summary for SVM model snap shot

4.3 Dividing the dataset into training and test set

The dataset used in this model building was 52,831 different vehicle types and 9 independent and 1 dependent attributes which have been saved as Comma Separated Version (CSV) format. In this experiment, dataset was divided into training set and testing set using 80:20 ratios. The training set was specifically used for the model building and the testing set was used for evaluation of the model.

Partitioning dataset in to training data and testing data using “createDataPartition” function in the ration 80/20 is depicted as follows.

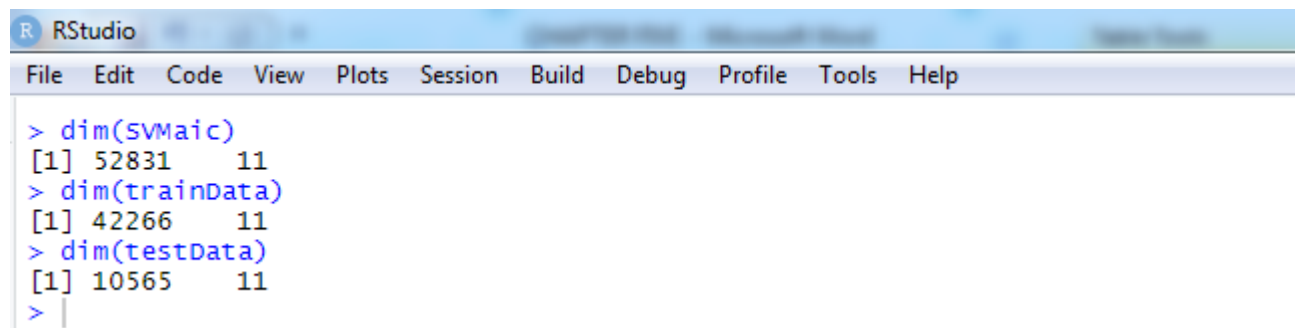
```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> MyPartition<- createDataPartition(y=SVMa1c$PRM,p=0.8,list = FALSE)
> trainData<- SVMa1c[MyPartition, ]
> testData<- SVMa1c[-MyPartition, ]
>

```

Figure 4. 4: Partitioning the dataset snap shot

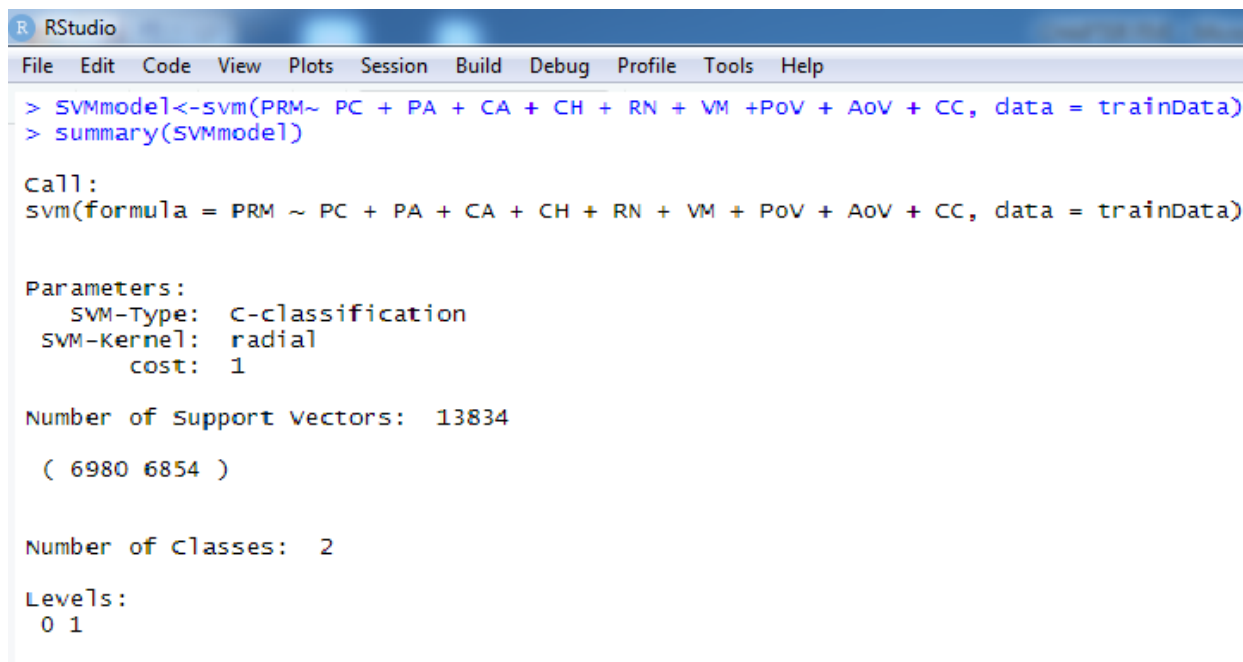
As indicated above in R programming, the dataset which is 52,831 was splitted into training set which is 80% (42, 265 records) and testing set which is 20% (10, 566 records)



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> dim(SVMa1c)
[1] 52831  11
> dim(trainData)
[1] 42266  11
> dim(testData)
[1] 10565  11
>
```

Figure 4. 5 : Dataset Partitioning Result for SVM model snap shot

Based on the data partitioning done above, we build model using the training data, then we test the model using the testing data.



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> SVMmodel<-svm(PRM~ PC + PA + CA + CH + RN + VM + PoV + AoV + CC, data = trainData)
> summary(SVMmodel)

Call:
svm(formula = PRM ~ PC + PA + CA + CH + RN + VM + PoV + AoV + CC, data = trainData)

Parameters:
  SVM-Type:  C-classification
 SVM-kernel: radial
      cost:  1

Number of Support Vectors: 13834
( 6980 6854 )

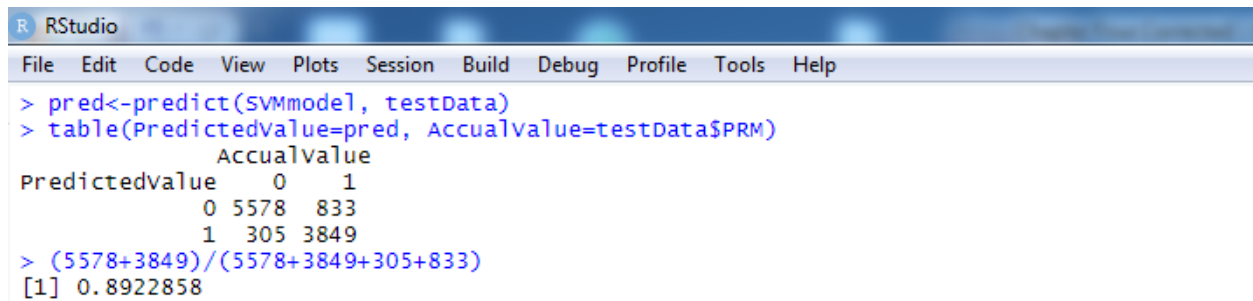
Number of Classes: 2

Levels:
0 1
```

Figure 4. 6 : Building SVM model snap shot

Note: The summary model with 13,834 support vectors where there were 6,980 vectors in the first class and 6,854 in the second class. These classes are 0 and 1 as depicted in the snap shot.

Testing the model accuracy using test data



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> pred<-predict(SVMmodel, testData)
> table(PredictedValue=pred, AccualValue=testData$PRM)
      AccualValue
PredictedValue 0 1
              0 5578 833
              1 305 3849
> (5578+3849)/(5578+3849+305+833)
[1] 0.8922858
```

Figure 4. 7 : Prediction and model accuracy for SVM model snap shot

Note: Figure 4. 7 shows the cross table which is presented as predicted value and accrual value. Based on the cross table above, the model performed as follows. The model predicted 5578 risk item records as risk items of standard rate or “Standard” which is actually “Standard” (TP), 305 predicted as “Loading” which is actually “Standard ” (FN), 833 predicted as “Standard” which is actually “Loading” (FP), 3849 predicted as “Loading” which is actually “Loading”(TN). Therefore the accuracy is calculated as $(TP+TN) / (TP+TN +FP+FN)$ which gives 89% of accurate. Moreover, more detail of the experiment can be seen with confusion Matrix

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
confusionMatrix(table(pred,testData$PRM))
Confusion Matrix and Statistics

      | Predicted \ Observed | 0      1 |
-----+-----+-----+-----
Observed 0 | 5578   833 |
          1 |  305  3849 |

              Accuracy : 0.8923
              95% CI   : (0.8862, 0.8981)
              No Information Rate : 0.5568
              P-Value [Acc > NIR] : < 2.2e-16

              Kappa   : 0.7792

              McNemar's Test P-value : < 2.2e-16

              Sensitivity : 0.9482
              Specificity : 0.8221
              Pos Pred Value : 0.8701
              Neg Pred Value : 0.9266
              Prevalence : 0.5568
              Detection Rate : 0.5280
              Detection Prevalence : 0.6068
              Balanced Accuracy : 0.8851

              'Positive' Class : 0

```

Figure 4. 8 : Confusion Matrix result for SVM model snap shot

The accuracy of SVM model as indicated in Figure 4.9 above is 94.82%. It is calculated as $(TP+TN) / (TP+TN+FP+FN) = (5578+3849) / (5578+3849+833+305) = 0.8923 = 89.23\%$. This implies the error rate or misclassification will be $(FN+FP) / \text{Total value} = (305+833) / (10565) = 0.1077$ or $1 - \text{Accuracy}$ which is $1 - 0.8923 = 0.1077$.

The precision of the model is when it predicts yes, how often it is correct? $(TP) / (TP+FP)$ which is $(5687) / (5687+61) = 98.94\%$.

The True Positive Rate also known as “Sensitivity” or “Recall” and it is to mean when it is actually yes, how often does the model predicts yes? Therefore, it is $(TP) / \text{Total yes}$. This is $(TP) / (TP+FN) = (5578) / (5578+833) = 0.8701 = 87.01\%$.

False positive rate is when it is actually No, how often the model predicts as Yes. Therefore it is computed as $(FP) / (\text{Total actual No}) = (833) / (833+3849) = 0.01779 = 1.779\%$

True Negative Rate is also known as “Specificity” is when it is actually No, how often the model predicts No? Therefore, it is $(TN) / (Total\ actual\ No) = (3849) / (833+3849) = 0.8221 = 82.21\%$

NB

4.4 Naïve Bayes Model

Naive Bayes algorithm has an ability of handle noisy data, continuous and discrete data and make probabilistic prediction and the main reason behind its popularity is that it can be written into the code very easily delivering predictions model in very less time. Thus, it can be used in the real-time model predictions (Huang and Lei, 2011)

Under this experiment building perdition model is done using Naive Bayes Classification algorithm and R programming

	PN	PC	PA	CA	CH	RN	VM	PoV	AoV	CC	PRM
1	AIC/GRJ/MTPV/307377/18	1	4400	0	0	1	8	2	1	7	0
2	AIC/GTR/MTPV/016692/13	1	8800	0	0	1	8	2	1	7	0
3	AIC/PIZ/MTPV/242860/17	1	6600	0	0	1	8	2	1	7	0
4	AIC/SAB/MTPV/271876/17	1	5205	0	0	1	8	2	1	7	0
5	AIC/KZC/MTPV/198101/16	1	8885	5000	1	1	8	2	2	7	1
6	AIC/KZC/MTPV/191774/16	1	27082	6900	1	1	8	2	1	9	1
7	AIC/ADD/MTPV/309060/18	1	13703	7000	1	1	8	2	1	9	1
8	AIC/BOL/MTPV/244734/17	1	8269	8000	1	1	8	2	2	7	1
9	AIC/BOL/MTPV/325165/18	1	10000	9200	1	1	2	2	1	6	1
10	AIC/FFN/MTPV/301691/18	1	10516	10000	1	1	8	2	1	9	1
11	AIC/GTR/MTPV/136671/15	1	13945	10350	1	1	7	2	1	6	1

Showing 1 to 14 of 52,831 entries, 11 total columns

Figure 4. 9 : Dataset snap shot for Naïve Bayes Model

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> str(NBaic)
'data.frame': 52831 obs. of 11 variables:
 $ PN : Factor w/ 21379 levels "AIC/22M/MTCM/027913/13",...: 12940 13195 18597 19105 152
48 15229 2042 6151 6388 10281 ...
 $ PC : int 1 1 1 1 1 1 1 1 1 1 ...
 $ PA : Factor w/ 14907 levels "1000","10000",...: 9738 14118 12272 10791 14173 6441 191
8 13737 2 353 ...
 $ CA : int 0 0 0 0 5000 6900 7000 8000 9200 10000 ...
 $ CH : int 0 0 0 0 1 1 1 1 1 1 ...
 $ RN : int 1 1 1 1 1 1 1 1 1 1 ...
 $ VM : int 8 8 8 8 8 8 8 8 2 8 ...
 $ PoV: int 2 2 2 2 2 2 2 2 2 2 ...
 $ AoV: int 1 1 1 1 2 1 1 2 1 1 ...
 $ CC : int 7 7 7 7 7 9 9 7 6 9 ...
 $ PRM: Factor w/ 2 levels "0","1": 1 1 1 1 2 2 2 2 2 2 ...
>
```

Figure 4. 10 : Data structure snap shot for Naïve Bayes Model

4.4.1 Data set splitting into training data and testing data

Similar to SVM the dataset is partitioned into training and testing dataset in the ration of 80:20 as depicted in Fig: 4.21 below.

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> mysplit_data<-sample.split(NBaic$PRM,splitRatio = 0.8)
> mytrain_data=subset(NBaic,mysplit_data==TRUE)
> mytest_data=subset(NBaic,mysplit_data==FALSE)
>
```

Figure 4. 11 : Dataset partitioning for Naïve Bayes model

4.4.2 Naïve Bayes model building

The models built using the training dataset and testing data will be used to check the accuracy and precision of the model as follows.

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> NBmodel=naiveBayes(x=mytrain_data, y=mytrain_data$PRM)
> prd=predict(NBmodel,newdata = mytest_data)
> table(ActualValue=mytest_data$PRM,predictedvalue=prd)
      predictedvalue
ActualValue  0    1
           0 5687 197
           1   61 4622
> (5687+4622)/(5687+4622+197+61)
[1] 0.9755844
> confusionMatrix(prd, mytest_data$PRM)
Confusion Matrix and Statistics

          Reference
Prediction  0    1
           0 5687  61
           1 197 4622

      Accuracy : 0.9756
      95% CI   : (0.9725, 0.9784)
 No Information Rate : 0.5568
 P-value [Acc > NIR] : < 2.2e-16

      Kappa   : 0.9507

 Mcnemar's Test P-value : < 2.2e-16

      Sensitivity : 0.9665
      Specificity : 0.9870
   Pos Pred Value : 0.9894
   Neg Pred Value : 0.9591
     Prevalence   : 0.5568
   Detection Rate : 0.5382
 Detection Prevalence : 0.5440
  Balanced Accuracy : 0.9767

      'Positive' class : 0
```

Figure 4. 12 : Model building and computation of accuracy for Naïve Bayes Model snap shot

The accuracy of Naïve Bayes model is 97.56% as depicted above in Figure 4.13. And therefore, the error rate or misclassification will be $(FN+FP)/ \text{Total value} = (61+192)/ (10562) = 0.024$ or 1-Accuracy which is $1-0.9756=0.024$.

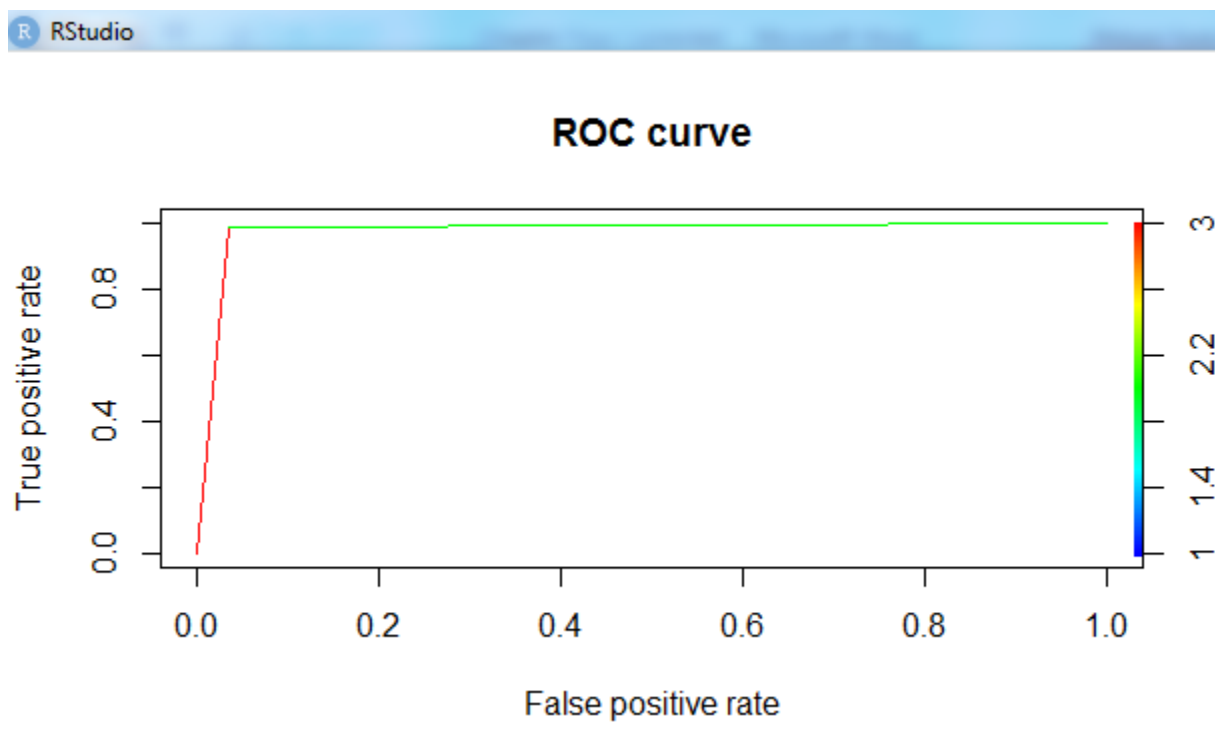
The precision of the model is when it predicts yes, how often it is correct? $(TP)/ (TP+FP)$ which is $(5687)/ (5687+61) = 98.94\%$.

The True Positive Rate also known as “Sensitivity” or “Recall” and it is to mean when it is actually yes, how often does the model predicts yes? Therefore, it is $(TP)/\text{Total yes}$. This is $(TP)/(TP+FN) = (5687)/(5687+197) = 0.9665 = 96.65\%$.

False positive rate is when it is actually No, how often the model predicts as Yes. This implies $(FP)/(\text{Total actual No}) = (61+4622) = 0.5873 = 58.73\%$

True Negative Rate (when the actual data is negative and the predictive model also predicted as negative. In this study when the actual data is 1 or loading which is negative sense and also the model is predicted the premium rate as 1 or loading) is also known as “Specificity” is when it is actually No, how often the model predicts No? Therefore, it is $(TN)/(\text{Total actual No}) = (4622)/(61+4622) = 0.9870 = 98.7\%$

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> prd2<-as.numeric(prd)
> RCORpre=prediction(prd2,mytest_data$PRM)
> RCORperf=performance(RCORpre, "tpr","fpr")
> plot(RCORperf, colorize=TRUE,main="ROC curve")
```



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> plot(RCORperf, colorize=TRUE,
+      main="ROC curve",
+      ylab="Sensitivity",
+      xlab="1-Specificity")
> auc<-performance(RCORpre,"auc")
> auc<-unlist(slot(auc,"y.values"))
> auc<-round(auc,4)
> legend(0.4,0.6, auc, title = "Area under ROC (AUC)", cex=1)
> |
```

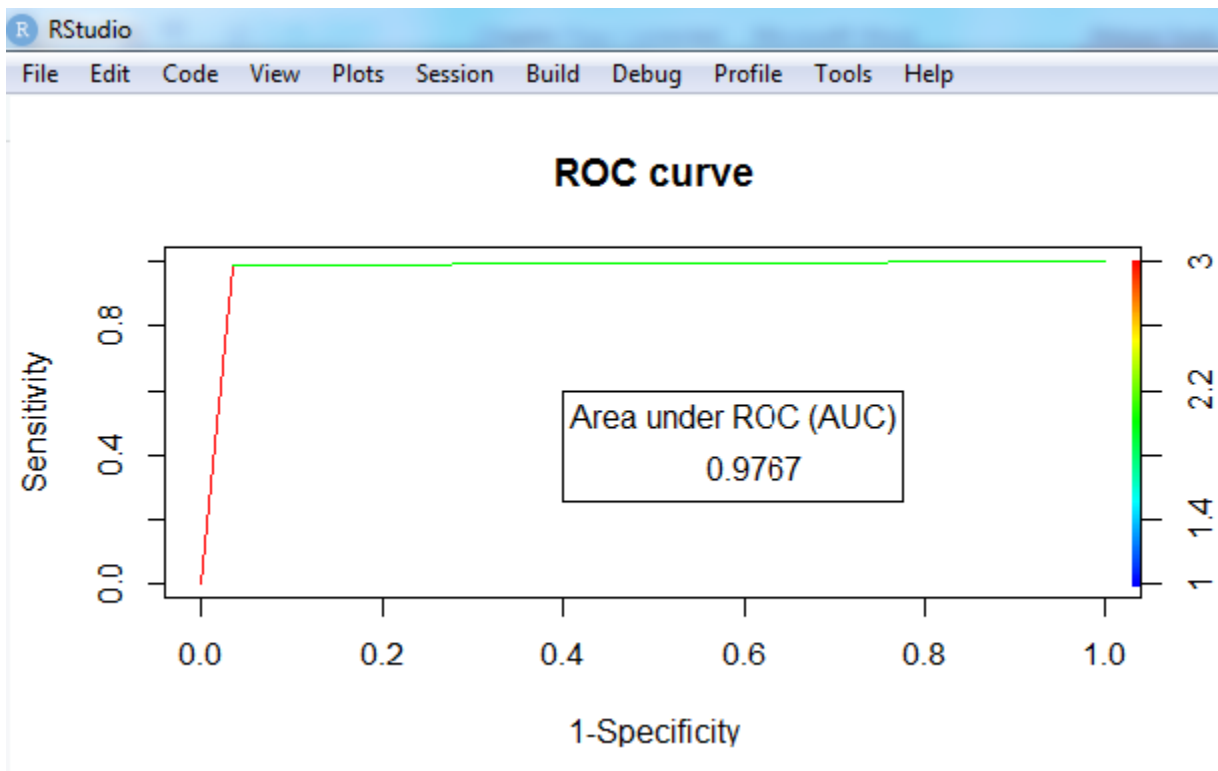


Figure 4. 13 : ROC Curve result snap shot for Naïve Bayes predictive model

As showed in figure 4.14 above the ROC is plotted between True Positive Rate (Sensitivity) or (Y axis) and False Positive Rate (1-Specificity) or (X Axis). In the above plot, the area under curve covers the maximum area (ROC curve 97.67%)

ROC determines the accuracy of a classification model and accuracy using Area Under Curve (AUC). The area under the curve (AUC) also referred to as the performance of the ROC curve.

4.5 Logistic Regression Modeling

There are three types of Logistic regression which can be used for prediction model (Linear Regression, Logistic Regression and Polynomial Regression). Linear regression is used in prediction model when there is a continuous relationship between dependent and independent variables, Logistic regression is a regression type in prediction model where the values of the dependent variable is categorical type like 0 or 1, Yes or No etc.. The other type of logistic regression is polynomial regression which is used for prediction model when the degree of independent variable is more than one. Under this experiment there are 9 independent variables and 1 dependent variable. The dependent variable Premium to be Calculated (PRM) has a categorical values which are either '0' or '1' which represents "Standard" and "Loading respectively". And therefore logistic regression is selected for the prediction model under this experiment.

Importing dataset to RStudio

A screenshot of the RStudio interface. The title bar shows 'RStudio'. The menu bar includes 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Help'. The console window contains the command:

```
> LRgaic<-read.csv(file.choose(),sep = ",")
```

Figure 4. 14 Importing dataset from csv file to RStudio for Logistic Regression snap shot

	PN	PC	PA	CA	CH	RN	VM	PoV	AoV	CC	PRM
1	AIC/GRJ/MTPV/307377/18	1	9738	0	0	1	8	2	1	7	0
2	AIC/GTR/MTPV/016692/13	1	14118	0	0	1	8	2	1	7	0
3	AIC/PIZ/MTPV/242860/17	1	12272	0	0	1	8	2	1	7	0
4	AIC/SAB/MTPV/271876/17	1	10791	0	0	1	8	2	1	7	0
5	AIC/KZC/MTPV/198101/16	1	14173	5000	1	1	8	2	2	7	1
6	AIC/KZC/MTPV/191774/16	1	6441	6900	1	1	8	2	1	9	1
7	AIC/ADD/MTPV/309060/18	1	1918	7000	1	1	8	2	1	9	1
8	AIC/BOL/MTPV/244734/17	1	13737	8000	1	1	8	2	2	7	1
9	AIC/BOL/MTPV/325165/18	1	2	9200	1	1	2	2	1	6	1
10	AIC/FFN/MTPV/301691/18	1	353	10000	1	1	8	2	1	9	1
11	AIC/GTR/MTPV/136671/15	1	2030	10350	1	1	7	2	1	6	1

Figure 4. 15 : Dataset for Logistic Regression Model

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> str(LRGaic)
'data.frame': 52831 obs. of 11 variables:
 $ PN : Factor w/ 21379 levels "AIC/22M/MTCM/027913/13",...: 12940 13195 18597 19105 15
48 15229 2042 6151 6388 10281 ...
 $ PC : num 1 1 1 1 1 1 1 1 1 1 ...
 $ PA : num 9738 14118 12272 10791 14173 ...
 $ CA : num 0 0 0 0 5000 6900 7000 8000 9200 10000 ...
 $ CH : num 0 0 0 0 1 1 1 1 1 1 ...
 $ RN : num 1 1 1 1 1 1 1 1 1 1 ...
 $ VM : num 8 8 8 8 8 8 8 8 2 8 ...
 $ PoV: num 2 2 2 2 2 2 2 2 2 2 ...
 $ AoV: num 1 1 1 1 2 1 1 2 1 1 ...
 $ CC : num 7 7 7 7 7 9 9 7 6 9 ...
 $ PRM: Factor w/ 2 levels "0","1": 1 1 1 1 2 2 2 2 2 2 ...

```

Figure 4. 16 : Dataset structure for Logistic regression model

Dataset splitting is done for this experiment in the same way of SVM and Naïve Bayes dataset splitting ratio which is 80:20 as depicted in Fig. 4.18 below.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> dataSplit<- sample(2,nrow(LRgaic),replace = TRUE,prob = c(0.8,0.2))
> training<- LRgaic[dataSplit==1,]
> testing<- LRgaic[dataSplit==2,]

```

Figure 4. 17 : Dataset partitioning for Logistic regression snap shot.

4.5.1 Logistic Regression model building

The model is built using the training data and its accuracy and precision is tested using the testing dataset as depicted below.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
> LRGmodel<-glm(PRM~ PC+ PA + CA + CH + RN + VM + PoV + AoV + CC,data=training,family
="binomial")
>
> pred<- predict(LRGmodel,testing,PRM="response")
> table(ActualValue=testing$PRM,predictedvalue=pred>0.5)
      predictedvalue
ActualValue FALSE TRUE
0          5439  472
1          2947 1729

RStudio
> Accuracy
> (5439+1729)/(5439+1729+2947+472)
[1] 0.6770568
> |

```

Figure 4. 18 : Logistic Regression Model building and prediction accuracy snap shot.

As it is depicted above in figure 4.19, the accuracy of the predictive model is 67.71%. The cross table above also shows that the predictive model correctly predicted 5439 records as risk items as risk items which deserve Standard rate where they are actually risk items on which standard rate are applied on (known as True Positive). On the other hand the model predicted 472 records as risk items with loading premium rate which are actually not (known as False Negative). The predictive model predicted 2947 records as Standard premium rating risk items which are actually

risk items with loading premium rating applied on (known as False Positive). Lastly, the model predicted 1729 records as risk items with loading premium rating where they are actually risk items with loading premium calculation (known as True Negative). Based on the above prediction results the accuracy of the model is calculated as $(TP+TN) / (TP+TN+FN+FP) = (5439+1729) / (5439+1729+472 +2947) = 0.6771 = 67.71\%$

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> summary(LRGmodel)

Call:
glm(formula = PRM ~ PC + PA + CA + CH + RN + VM + PoV + AoV +
     CC, family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2244 -0.9482 -0.6393  1.0599  2.5616

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.108e+00  1.076e-01  38.197 < 2e-16 ***
PC           4.654e-01  2.985e-02  15.595 < 2e-16 ***
PA          -1.522e-05  2.870e-06  -5.304 1.13e-07 ***
CA           8.410e-07  1.153e-07   7.295 2.98e-13 ***
CH          -4.765e-01  2.556e-02 -18.644 < 2e-16 ***
RN          -4.679e-02  2.046e-03 -22.864 < 2e-16 ***
VM           3.529e-02  7.072e-03   4.990 6.04e-07 ***
PoV         -3.161e-01  9.573e-03 -33.021 < 2e-16 ***
AoV          4.193e-01  8.548e-03  49.051 < 2e-16 ***
CC          -1.770e-01  5.250e-03 -33.709 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58142  on 42352  degrees of freedom
Residual deviance: 51210  on 42343  degrees of freedom
AIC: 51230

Number of Fisher Scoring iterations: 4
>

```

Figure 4. 19 : The output of logistic regression model snap shot

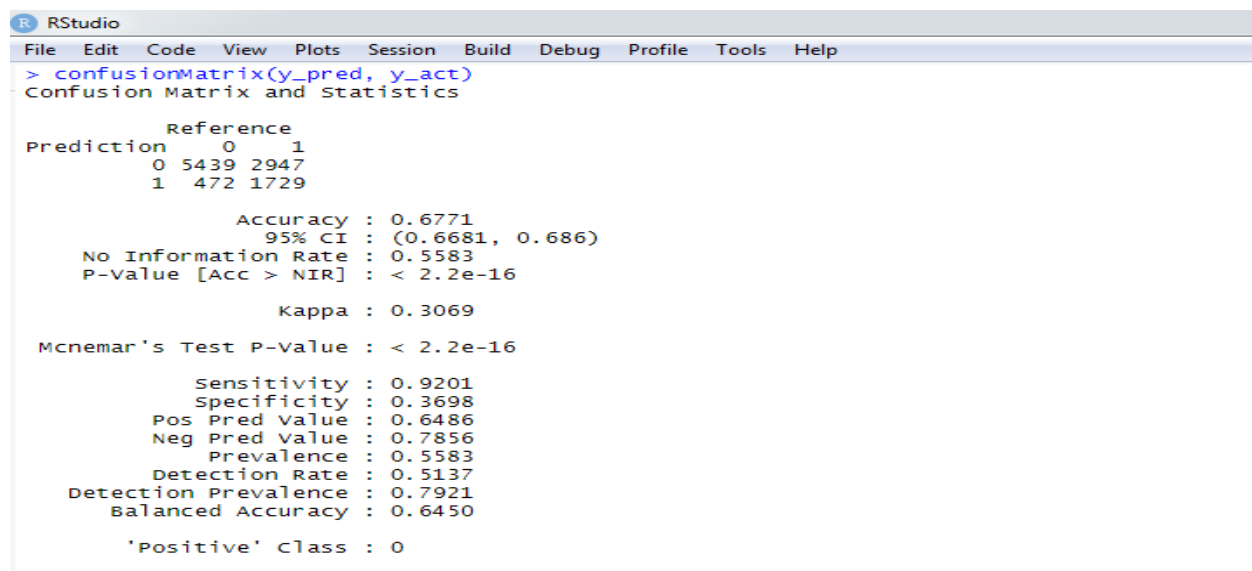
The figure. 4.20 above shows the values of summary of the model. The Estimate values are coefficient values for each independent variable. The other important point is the stars (***) corresponding to each independent variable. They show the importance of each variable as indicated. If there is no star or dot (.) value for a variable under experiment, it shows it is not significant for the model and can be removed from the experiment. In this experiment all the independent variables show 99.9 confident that it is significant for the model.

Null deviance: the deviance from actual dataset when we use only Intercept removing other independent variables. Therefore, the null deviance of this experiment removing all the independent variable is 58142. Therefore, null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean).

Residual deviance: is the deviance when we use independent variables which obviously minimize the value when compared to null deviance, in this experiment the residual deviance is 51210 which is less than 58142 (values of Null deviance). Therefore, Residual deviance shows how well the response variable is predicted with inclusion of independent variables.

AIC: is a value of an experiment that should be at minimum as much as possible. The value of AIC will be optimal when non-significant variable are removed from the experiment.

The confusion matrix show detail information of the logistic regression model as depicted in Fig. 4.21 below.



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> confusionMatrix(y_pred, y_act)
Confusion Matrix and Statistics

      Reference
Prediction 0  1
 0  5439 2947
 1   472 1729

      Accuracy : 0.6771
      95% CI   : (0.6681, 0.686)
 No Information Rate : 0.5583
 P-value [Acc > NIR] : < 2.2e-16

      kappa : 0.3069

McNemar's Test P-Value : < 2.2e-16

      Sensitivity : 0.9201
      Specificity : 0.3698
 Pos Pred Value : 0.6486
 Neg Pred Value : 0.7856
 Prevalence : 0.5583
 Detection Rate : 0.5137
 Detection Prevalence : 0.7921
 Balanced Accuracy : 0.6450

      'Positive' Class : 0
```

Figure 4. 20 : Confusion Matrix for Logistic Regression snap shot.

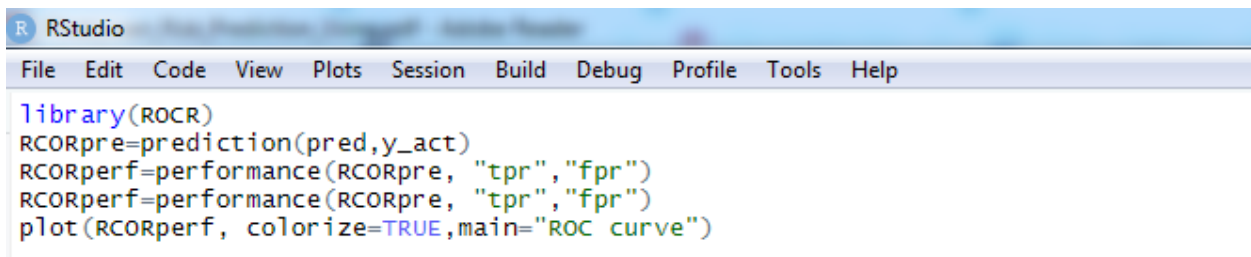
The accuracy of logistic regression in figure 4.21 above is 67.71%. This implies that the error rate or misclassification will be $(FN+FP)/\text{Total value} = (2947+472)/(10587) = 0.3229$ or $1 - \text{Accuracy}$ which is $1 - 0.6771 = 0.3229$ and the.

Precision of the model is when it predicts yes, how often it is correct? $(TP)/(TP+FP)$ which is $(5439)/(5439+2947) = 64.86\%$.

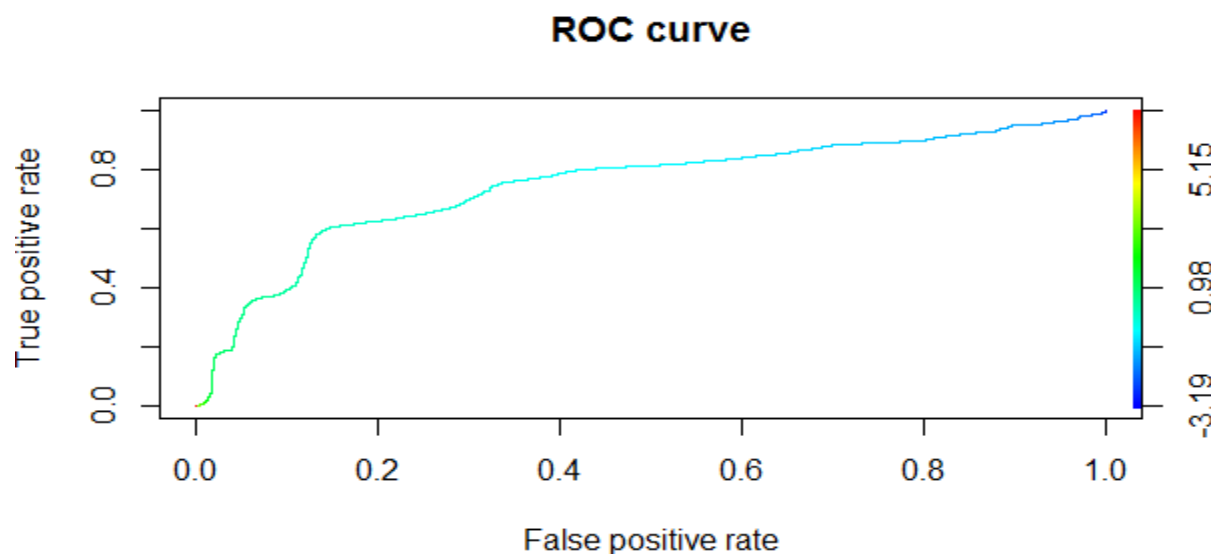
The other point in the output of fig 4.29 is True Positive Rate. Sensitivity is also known as “Sensitivity” or “Recall” and it is to mean when it is actually yes, how often does the model predicts yes? Therefore, it is $(TP)/\text{Total yes}$. This implies $(TP)/(TP+FN) = (5439)/(5439+742) = 0.92 = 92\%$.

False positive rate is when it is actually No, how often the model predicts as Yes. This implies $(FP)/(\text{Total actual No}) = (2947)/(2947+1729) = 0.6302 = 63.02\%$

True Negative Rate is also known as “Specificity” is when it is actually No, how often the model predicts No? Therefore, it is $(TN)/(\text{Total actual No}) = (1729)/(2947+1729) = 0.36976 = 36.98\%$



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
library(ROCR)
RCORpre=prediction(pred,y_act)
RCORperf=performance(RCORpre, "tpr","fpr")
RCORperf=performance(RCORpre, "tpr","fpr")
plot(RCORperf, colorize=TRUE,main="ROC curve")
```



```
plot(RCORperf, colorize=TRUE,  
     main="ROC curve",  
     ylab="Sensitivity",  
     xlab="1-Specificity")  
auc<-performance(RCORpre,"auc")  
auc<-unlist(slot(auc,"y.values"))  
auc<-round(auc,4)  
legend(0.4,0.6, auc, title = "Area under ROC (AUC)", cex=1)
```

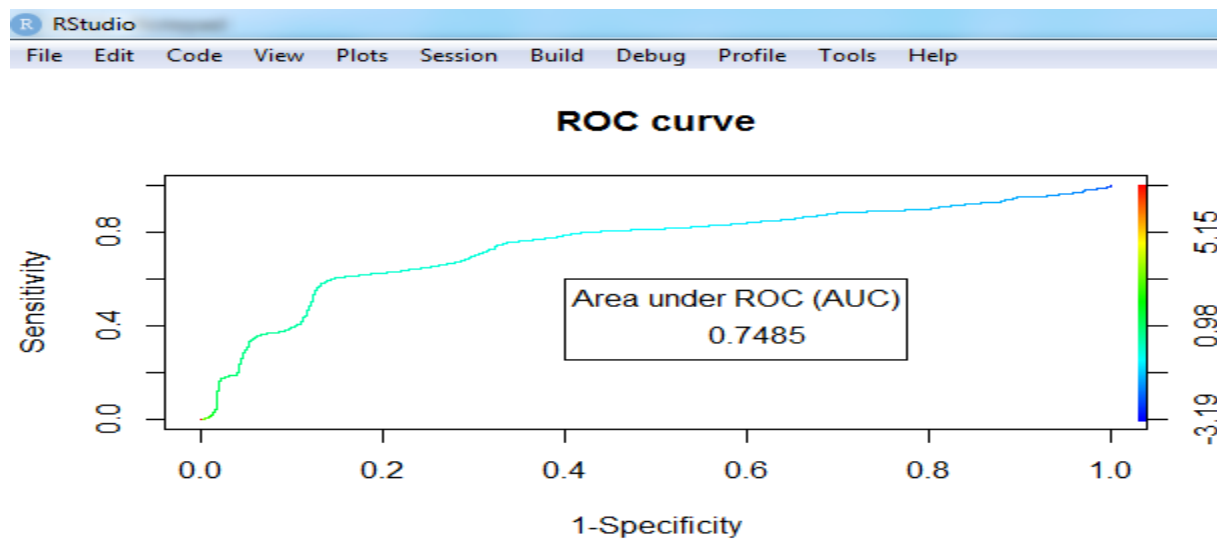


Figure 4. 21: ROC Curve result snap shot

As showed in fig.4.22 above the ROC is plotted between True Positive Rate (Sensitivity) or (Y axis) and False Positive Rate (1-Specificity) or (X Axis). In the above plot, the area under curve covers the maximum area (ROC curve 74.85%)

ROC determines the accuracy of a classification model at a user defined threshold value. It determines the model's accuracy using Area Under Curve (AUC). The area under the curve (AUC) also referred to as the performance of the ROC curve.

4.6 Summary of the findings

The current claim cost control trend in (Awash insurance S.C) is inefficient due to its manual controlling mechanism. According to the discussion made with Underwriting and Branch operations manager, every branch reports the claim happened monthly. The vehicles engaged into the claim will be gathering from every branch and compiled at head office and then served to the management table for decision. Most of the time the decision will be premium rate revision. Here the problem is, first it is too difficult to address manually all the cases exhaustively, second the current way of analysis doesn't take into consideration the concept of parameters for each and every risk item class.

Taking into account the current problems observed, an experiment was extensively conducted on the dataset using different prediction models and different outcomes observed based on the prediction model (SVM, Naïve Bayes and Logistic Regression). These predictive models are developed using selected attributed of risk item under motor class of business after exhaustive discussion with underwriters, branch managers and claim offices as well as claim managers.

The objective of this study is to identify under which premium calculation risk items could be categorize (Standard which is normal company rate and Loading which include charging additional premium).And the experiments performed mainly to identify the best classifier for predicting the risk item to the correct category. The following comparison is made to select the best classifier that predicts the right category of the risk items.

Table 3. 19: Comparison of prediction models

Classifier	Accuracy	precision
SVM	89.81%	82.21%
Naïve Bayes	97.56%	98.7%
Logistic Regression	67.33%	36.77%

The percentage of correctly classified instances is referred as model accuracy and the accuracy is referred to as model performance (Ponnuraja, Lakshmanan and Srinivasan, 2016). Therefore, as it is observed from table 3.19 above, Naïve Bayes performed better when compared to SVM and Logistic Regression in both accuracy and precision while Logistic Regression performed lower.

Regarding the research question: *How best can data mining technique address the existing challenges?* The best performing prediction model will be employed and therefore, during underwriting process a user can predict the claim cost of risk item and categorize under appropriate premium calculation group (Standard rate application or Loading application group) and serve the customer using Naïve Bayes prediction model.

The variable used under this experiment are very important as depicted in Figure 4.20, that means the level of significance of each variable is indicated as three stars which indicates the highest significance of each variable. So using these variable the users at underwriting stage can capture values of each variable and predict the claim cost and take appropriate measure with the consent of the branch manager.

The end users of Awash insurance Company do not need to know the detail of the application at the back of the system. In order to deploy this experiment, the script used under this experiment has to be integrated to some programming languages and connoted to the database of Awash insurance company which is central database server. At the time the user click on the “Premium to be Calculated” button on the user Interface, the click event triggers the script integrated with the programming language. Then it consults the database and give suggested Premium calculation rate. To save performance of the system and as the database of the company is dynamically changed from time to time the dataset partitioning process can be scheduled on demand which can

support offline prediction too. Moreover, the system suggestion should not be mandatory; the final decision is the management decision as the insurance business is very flexible. This actual deployment process is the future work of the experiment. The user interface might look like the following. But it doesn't mean the script at the back only contains this part. This is the simplified one.

4.7 Deployment

4.7.1 Prototype

The deployment of this experiment will be at the underwriting directorate of the company to be used by each and every branch during the underwriting process. As the building of a model takes much time, the model should be built once and saved to some directory to be used anytime in demand. When a customer brings a motor vehicle to Awash insurance company, to be insured, the user captures required parameters. If the number of vehicles is one (Table 4.1), the required parameters will be captured as depicted in Figure 4.25 below. The captured data should be framed in line with the built model format as depicted in Figure 4.26. After capturing and framing the required fields identified during the experiment, the saved model should be loaded into R studio and used as depicted in Figure 4.27. Finally, the new data captured will be used to predict the upcoming claim cost as indicated in Figure 4.28.

Table 4.1: Required parameters for a single vehicle.

Parameters	Represented by	Values	Represented by
Policy Name	PN	AIC/BOL/MTPV/203020/20	AIC/BOL/MTPV/203020/20
Policy Code	PC	MTPV	1
Premium Amount	PA	5000	5000
Claim Amount	CA	1500	1500
Claim History	CH	Yes	1
Risk Item Name	RN	Isuzu Truck FSR	32
Vehicle Make	VM	JAPAN	8
Purpose of the Vehicle	PoV	General Cartage	5
Age of the Vehicle	AoV	12, (Age group 11-15)	3
Carrying Capacity	CC	2389, (CC group 2351-3050)	7

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> MyPartition<- createDataPartition(y=SVMaic$PRM,p=0.8,list = FALSE)
> trainData<- SVMaic[MyPartition, ]
> testData<- SVMaic[-MyPartition, ]
>
> SVMmodel<-svm(PRM~ PC + PA + CA + CH + RN + VM +PoV + AoV + CC, data = trainData)
> SVMmodel

Call:
svm(formula = PRM ~ PC + PA + CA + CH + RN + VM + PoV + AoV + CC, data = trainData)

```

Figure 4. 22 : Building a model

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> save(SVMmodel,file = "D:/PredictiveModels/SVMmodel.rda")
>

```

Figure 4. 23: Saving a mode to specific directory

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> library(e1071)
> PN<-c("AIC/BOL/MTPV/203020/20")
> PC<-c(1)
> PA<-c(5000)
> CA<-c(1500)
> CH<-c(1)
> RN<-c(32)
> VM<-c(8)
> PoV<-c(5)
> AoV<-c(3)
> CC<-c(7)

```

Figure 4. 24 : Capturing vehicle’s important parameters

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> df<-data.frame(PN,PC,PA,CA,CH,RN,VM,PoV,AoV,CC)
> df

```

	PN	PC	PA	CA	CH	RN	VM	PoV	AoV	CC
1	AIC/BOL/MTPV/203020/20	1	5000	1500	1	32	8	5	3	7

Figure 4. 25 : Putting the captured data into appropriate model format.

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> load("D:\PredictiveModels\SVMmodel.rda")
> SVMmodel

Call:
svm(formula = PRM ~ PC + PA + CA + CH + RN + VM + PoV + AoV + CC, data = trainData)

```

Figure 4. 26: Loading the saved model to R studio

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> pred2<-predict(SVMmodel,newdata = df)
> pred2
1
1
Levels: 0 1
>

```

Figure 4. 27 : Making prediction using the newly captured data

Note: the rate user for premium calculation (PRM) has categorical values, either ‘0’ or ‘1’ representing “Standard” and “Loading” respectively. The prediction result shows PRM value is 1 for record number 1. It means this vehicle’s claim cost is high. Therefore, based on the prediction result, the user is advised to load the premium to give insurance cover policy. Moreover, the customer may bring to the insurance company more than one vehicle. For instance, the number of vehicles can be 5. Under that situation, the predictive model can be used as follows

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> # when Five vehicles come to Awash insurance company
>
> library(e1071)
> PN<-c("AIC/BOL/MTPV/203020/20", "AIC/ADD/MTCM/023499/13", "AIC/GRJ/MTCM/034600/13", "AIC/BHD/MTCM/2016
81/16",
+ "AIC/FFN/MTPV/038159/13")
> PC<-c(1,0,0,0,1)
> PA<-c(5000,5112,6696,8013,8239)
> CA<-c(1500,0,5000,200951,31004)
> CH<-c(1,0,1,1,1)
> RN<-c(32,2,26,15,46)
> VM<-c(8,8,7,8,8)
> PoV<-c(5,11,5,5,11)
> AoV<-c(3,2,2,2,6)
> CC<-c(7,7,10,10,1)

```

Figure 4. 28 : Capturing vehicle’s information (required Parameters)

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> df<-data.frame(PN,PC,PA,CA,CH,RN,VM,PoV,AoV,CC)
> df
  
```

		PN	PC	PA	CA	CH	RN	VM	PoV	AoV	CC
1	AIC/BOL/MTPV/203020/20	1	5000	1500	1	32	8	5	3	7	
2	AIC/ADD/MTCM/023499/13	0	5112		0	0	2	8	11	2	7
3	AIC/GRJ/MTCM/034600/13	0	6696	5000	1	26	7	5	2	10	
4	AIC/BHD/MTCM/201681/16	0	8013	200951	1	15	8	5	2	10	
5	AIC/FFN/MTPV/038159/13	1	8239	31004	1	46	8	11	6	1	

Figure 4. 29: putting the captured data into appropriate model format.

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> load("D:\PredictiveModels/SVMmodel.rda")
> SVMmodel

Call:
svm(formula = PRM ~ PC + PA + CA + CH + RN + VM + PoV + AoV + CC, data = trainData)

```

Figure 4. 30: loading the saved model to R studio

```

R RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
> pred2<-predict(SVMmodel,newdata = df)
> pred2
1 2 3 4 5
1 0 0 1 1
Levels: 0 1

```

Figure 4. 31: predictions using the newly captured data

Note: the user captured identified attributes of the 5 vehicles based the parameters captured, the model, predicted the premium to be collected from vehicle number 1, 4 and 5 should be loaded and premium from vehicle number 2 and 3 should be calculated using standard rate of the company.

CHAPTER FIVE

5 CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter summarizes and concludes findings of the work done based on the findings of the study. The conclusion and recommendations as well as future work will be discussed briefly as follows.

5.2 Conclusion

The purpose of this study was to determine how machine learning model could be used to assess and identify the risk item's claim cost that they can incur in future during underwriting if they are provided an insurance coverage.

Awash insurance company already automated its operation eight years back which means all operation and non-operation actives are managed centrally at head office. All the transaction performed anywhere in the company are actually performed in the central database at head office. The dataset used for the purpose of this study is collected from the company's database or central database server. On underwriting module or proposal page, there are many fields which are not mandatory that a user easily skip without filling them during underwriting which create incomplete data record. As the result there were many records found which are incomplete of required parameter in this experiment. Based on the branch they are originated, extensive interaction has been made with the branch underwriters and manager to complete them. Those records whose policy periods are too long back and enough information couldn't be found are removed from the experiment. Finally 52,831 complete record of dataset were brought to the experiment.

From the many parameter of underwriting and claim data record, this experiment employed 9 parameters. These are policy code (PC) - the code that identifies where the vehicle is motor commercial or motor private, Premium collected (PA) - the premium amount collected from the risk item, Claim Amount (CA) - the claim amount paid for that risk item, Claim history (CH) -

where that particular risk item experienced claim or not, Risk Item Name (RN) - the name of the risk item, Vehicle make (VM) - the country in which the vehicle is manufactured, Purpose of the Vehicle (PoV) - which differentiate for what purpose the vehicle is used for, Age of the vehicle (AoV) - the age of the vehicle since manufactured and Carrying Capacity (CC) - the carrying capacity of the vehicle. The importance of these parameters is experimented and the output for all of them is very significant.

The experiment has been conducted on the dataset using different perdition models namely SVM, Naïve Bayes and Logistic Regression. All the prediction models were resulted in different accuracy levels. The experiment outcome for each model – SVM, Naïve Bayes and Logistic Regression is 89.81%, 97.56%, 67.33% respectively. Based on the experiment outcome of each prediction model, Naïve Bayes performed better with prediction accuracy of 97.56% followed by SVM with prediction accuracy of 89.81% and the least according to the experiment outcome was Logistic regression which predicted correctly 67.33%.

Based on the experiments conducted and output found it can be concluded that the Machine learning techniques can be used in an insurance industry to categorize risk items according to their risk exposure using predictive models by measuring their performance. Therefore, other insurance companies can also use predictive models to classify the risk items according to claim exposure of the risk items automatically during underwriting which increase their profitability as they have a chance to adjust their premium ahead of insuring the risk.

5.3 Recommendations and future work

The experiment conducted has shown an encouraging result, but it doesn't mean it can be put in place at user's station easily. There might be different factors which need further investigation that requires technically rich professionals. Therefore, other researchers and academicians who are interested in the same area can explore in detail and come up with better deployment method to take this predictive model into practice.

The researcher forwards the following points as a future work

- The experiment conducted in this research paper may be used as an incentive for further investigation so that it can be implemented in insurance companies. In order to deploy this finding practically in insurance companies working environment, researchers interested in the same area needs to integrate the scripts used in this experiment to some programming language and able to manipulate the database at the back.
- As the end users are not interested into detail of the process to check the probability of claim for each risk item, they need simplified user interface. Therefore, interested researchers need to investigate an optimal user interface (dashboard) so that single click event can trigger the integrated script to manipulate the database for prediction purpose.
- The prediction process needs CPU time that might make the production environment busy as the number of users increasing from time to time. Therefore, it requires further investigation to use the database of the company's offline.

References

- Abdul Alhassan and Nicholas Biekpe (2015), “Determinants of Life Insurance Consumption in Africa”, Research in International Business and Finance.
- Burri, Bojja and Buruga “Insurance Claim Analysis Using Machine Learning Algorithms”. International Journal of Innovative Technology and Exploring Engineering, Volume-8, April 2019
- Daniel Mehari and Tilahun Aemiro (2013), “Firm Specific Factors that Determine Insurance Companies’ Performance in Ethiopia”, European Scientific Journal, Vol.9, No 10, PP. 1857-7881.
- Dawei, J. (2011), “The Application of Data Mining in Knowledge Management”, International Conference on Management of e-Commerce and e-Government, IEEE Computer Society.
- Demis Hailegebreal (2016), “Macroeconomic and Firm Specific Determinants of Profitability of Insurance Industry in Ethiopia”, Global Journal of Management and Business Research, Vol. 16, No 7
- Deshpande and Thakare (2010), “DATA MINING SYSTEM AND APPLICATIONS: A REVIEW”, International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010.
- Sheaba Rani and Sekata Gonena, (2017) “Comparative study on motor insurance practices of public and private insurance companies focusing on customer satisfaction”, International Journal of Commerce and Management Research, V.3, No. 3, pp 41-47
- D.O. Olayungbo (2015). “Effects of Life and Non-Life Insurance on Economic Growth in Nigeria: An Autoregressive Distributed Lag (ARDL) Approach” , Global Journal of Management and Business Research ,Vol. 15, No.11

- DR. SAMBASIVAM and MR. ABATE GASHAW, “A STUDY ON THE PERFORMANCE OF INSURANCE COMPANIES IN ETHIOPIA”, International Journal of Marketing, Financial Services & Management Research, Vol.2, No. 7, pp. 2277- 3622
- François (2016) retrieved from <https://www.researchgate.net/publication/302305568>
- Gaurav Akrani (2011), retrieved from <https://kalyan-city.blogspot.com/2011/03/principle-of-insurance-7-basic-general.html>
- Graham J. Williams and Zhexue Huang (2013) , “Dependent frequency–severity modeling of insurance claims”, Insurance: Mathematics and Economics Vol 64, pp. 417–428
- Hailmichael (2011), “The Ethiopian Insurance Market”, African Reinsurance Corporation
- Hailu, Z, 1., ed. (2007), “Insurance In Ethiopia: historical development, present status and future challenges”, 1 edition, Addis Ababa, Ethiopia.
- Hanafizadeh and Paydar (2013), “A Data Mining Model for Risk Assessment and Customer Segmentation in the Insurance Industry”, International Journal of Strategic Decision Sciences, 4(1), PP. 52-78.
- Hifza Malik (2011). “DETERMINANTS OF INSURANCE COMPANIES PROFITABILITY:AN ANALYSIS OF INSURANCE SECTOR OF PAKISTAN”, Academic Research International, V.1, No 3
- Jayanthi Ranjan (2009) ” Data mining in pharma sector : benefits”, IJHCQA, Vol. 22 No. 1, pp. 82-92.
- Jihye Jeon (, 2015) “the Strengths and Limitations of the Statistical Modeling of Complex Social Phenomenon: Focusing on SEM, Path Analysis, or Multiple Regression Models”, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering Vol:9, No:5

- Kaviani and Dhotre (2017), “Short Survey on Naive Bayes Algorithm”, International Journal of Advance Research in Computer Science and Management,
- K.P.M.L.P. Weerasinghe and M.C. Wijegunasekara (2016), “Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims” European International Journal of Science and Technology Vol. 5
- K.P.M.L.P. Weerasinghe and M.C. Wijegunasekara, “A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims”, European International Journal of Science and Technology, Vol. 5, No. 1
- K.Umamaheswari and DR. S.Janakiraman (2014). “Role of Data mining in Insurance Industry”, International Journal of Advanced Computer Technology, Vol. 3, No. 6
- Muluken Alemu, (2015). “Application Of Data Mining Techniques For Student Success And Failure Prediction “ (The Case Of Debre_Markos University), INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VO. 4, ISSUE 04
- Nadali, kakhky and Nosratabadi (2016). “Evaluating the Success Level of Data Mining Projects Based on CRISP-DM Methodology”
<https://www.researchgate.net/publication/261109414>
- Naveed Ahmed, Zulfqar Ahmed and Ishfaq Ahmed, (2010) “Determinants of Capital Structure: A Case of Life Insurance Sector of Pakistan”, European Journal of Economics, Finance and Administrative Sciences, No. 24
- Persson (2008), “Road traffic accidents in Ethiopia”, Magnitude, causes and possible interventions
- Ponnuraja, Lakshmanan and Srinivasan (2016), “Performance Accuracy between Classifiers in Sustain of Disease Conversion for Clinical Trial Tuberculosis Data: Data Mining Approach”, IOSR Journal of Dental and Medical Sciences (IOSR-JDMS), Vol. 15, No.4.

- RavindraChangala, D.Rajeswara Rao, T.Janardhana Rao, P.Kiran Kumar and Kareemunnisa, “Knowledge Discovery Process: The Next Step for Knowledge Search”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, No. 5.
- Sadiku ,Shadare, Musa and Akujuobi (2016) “DATA VISUALIZATION”, International Journal of Engineering Research And Advanced Technology(IJERAT),Vol.2,No.12.
- Sahu,Shrma and Gondhalakar (2013), “A Brief Overview on Data Mining Survey”, International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volum1, Issue 3
- Santra and Christy (2012) “Genetic Algorithm and Confusion Matrix for Document Clustering”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2.
- Sastry and Babu, (2013). Implementation of CRISPMethodology forERP Systems, International Journal of Computer Science Engineering (IJCSE), Vol. 2 No.05.
- Shafique and Qaiser (2014) , “A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA”, International Journal of Innovation and Scientific Research , ISSN 2351-8014 ,Vol. 12 No. 1 pp. 217-222
- Sibindi (2015), “ THE ART OF ALTERNATIVE RISK TRANSFER METHODS OF INSURANCE”, Risk governance & control: financial markets & institutions / Volume 5, Issue 4.
- Silwattananusarn and Tuamsuk (2012), “Data Mining and Its Applications for Knowledge Management “, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5.
- Singh and Kumar (2012). “Conceptual Mapping of Insurance Risk Management to Data Mining”, International Journal of Computer Applications (0975 – 8887) Volume 39– No.2.

- Srivastava and Bhambhu (2010) “Data classification using support vector machine”, Journal of Theoretical and Applied Information Technology,
- Sumathi, Kannan & Nagarajan, (2016), “Data Mining: Analysis of student database using Classification Techniques” International Journal of Computer Applications (0975 – 8887) Volume 141 – No.8.
- Tejashree and Sawant (2016) “R: Data Mining Tool And Its Applications”, International Journal of Advanced Computer Technology & Management (IJACTM) , Vol 1,ISSN : 2343-662X.
- Temesgen Zeleke (2004), “Motor Risks and the Current Status of Motor insurance in Ethiopia”, Birritu, No.90 August-October, pp 19-38, Addis Ababa
- Teklit Berhe and Prof. Jasmindeep Kaur (2017), “Determinants of insurance companies’ profitability Analysis of insurance sector in Ethiopia”, International Journal of Research in Finance and Marketing (IJRFM), Vol. 7 ,No. 4, pp. 124~137
- Umamaheswari Janakiraman (2014) “Role of Data mining in Insurance Industry”, an international journal of advanced computer technology, 3 (6), June-2014 (Volume-III, Issue-VI)
- U. Shafique and H. Qaiser (2014) , “A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)” International Journal of Innovation and Scientific Research, vol. 12, pp. 217-222.
- Yu Yan, HaiyingXie (2009), “Research on the Application of Data Mining Technology in Insurance In formalization” In Proceedings of the International conference on Hybrid Intelligent Systems (HIS), PP. 202-205 © IEEE.
- Yuguang, Huang and Lei Li (2011.) ” Naïve Bayes Classification Algorithm Based on Small Sample Set” Beijing University. of Posts and Telecommunications, Proceedings of IEEE CCIS

- Zuriahati Yunos, Ali yuddin, and Noriszura Ismail (2016). “Predictive Modelling for Motor Insurance Claims Using Artificial Neural Networks”, International Journal of Advances in Soft Computing and its Applications,

Annex 1

Script for Naïve Bayes Model development

```
install.packages("MASS")
install.packages("rpart")
install.packages("caTools")
install.packages("naivebayes")
install.packages("scales")
install.packages('lattice')
install.packages("ggplot2")
install.packages("caret")
install.packages('e1071')
library(scales)
library(naivebayes)
library(MASS)
library(rpart)
library(caTools)
library(naivebayes)
library(e1071)
library(lattice)
library(ggplot2)
library(caret)
set.seed(1234)
Nbaic<-read.csv(file.choose(),sep = ",")
str(Nbaic)
Nbaic$PRM<-as.factor(Nbaic$PRM)
str(Nbaic)
Nbaic$PRM<- factor(Nbaic$PRM,levels=c(0,1))
mysplit_data<-sample.split(Nbaic$PRM,SplitRatio = 0.8)
mytrain_data=subset(Nbaic,mysplit_data==TRUE)
mytest_data=subset(Nbaic,mysplit_data==FALSE)
```

```
NBmodel=naiveBayes(x=mytrain_data, y=mytrain_data$PRM)
prd=predict(NBmodel,newdata = mytest_data)
table(ActualValue=mytest_data$PRM,predictedvalue=prd)
(5687+4622)/(5687+4622+197+61)
confusionMatrix(prd, mytest_data$PRM)
```

Annex 2

Script for Logistic Regression prototype

```
PN<-c("AIC/GTR/MTPV/136676/15")
PC<-c(1)
PA<-c(1465)
CA<-c(12000)
CH<-c(1)
RN<-c(1)
VM<-c(7)
PoV<-c(2)
AoV<-c(2)
CC<-c(9)
df<-data.frame(PN,PC,PA,CA,CH,RN,VM,PoV,AoV,CC)
df
pred2<-predict(LRGmodel,newdata = df)
pred2
PN<-c("AIC/GTR/MTPV/136676/15","AIC/ADD/MTCM/023499/13")
PC<-c(1,0)
PA<-c(1465,5112)
CA<-c(12000,0)
CH<-c(1,0)
RN<-c(1,2)
VM<-c(7,8)
PoV<-c(2,11)
AoV<-c(2,2)
CC<-c(9,7)
df<-data.frame(PN,PC,PA,CA,CH,RN,VM,PoV,AoV,CC)
df
pred2<-predict(LRGmodel,newdata = df)
pred2
PN<-
c("AIC/GTR/MTPV/136676/15","AIC/ADD/MTCM/023499/13","AIC/KZC/MTPV/290406/18
")
PC<-c(1,0,1)
PA<-c(1465,5112,10968)
CA<-c(12000,0,38323)
CH<-c(1,0,1)
RN<-c(1,2,1)
```



```

VM<-c(7,8,8)
PoV<-c(2,11,2)
AoV<-c(2,2,1)
CC<-c(9,7,9)
df<-data.frame(PN,PC,PA,CA,CH,RN,VM,PoV,AoV,CC)
df
pred2<-predict(LRGmodel,newdata = df)
pred2
PN<-
c("AIC/GTR/MTPV/136676/15","AIC/ADD/MTCM/023499/13","AIC/KZC/MTPV/290406/18
","AIC/ADM/MTCM/037243/13")
PC<-c(1,0,1,0)
PA<-c(1465,5112,10968,8867)
CA<-c(12000,0,38323,30000)
CH<-c(1,0,1,1)
RN<-c(1,2,1,2)
VM<-c(7,8,8,8)
PoV<-c(2,11,2,11)
AoV<-c(2,2,1,2)
CC<-c(9,7,9,9)
df<-data.frame(PN,PC,PA,CA,CH,RN,VM,PoV,AoV,CC)
df
pred2<-predict(LRGmodel,newdata = df)
pred2
PN<-
c("AIC/GTR/MTPV/136676/15","AIC/ADD/MTCM/023499/13","AIC/KZC/MTPV/290406/18
","AIC/ADM/MTCM/037243/13","AIC/AMB/MTCM/089642/14")
PC<-c(1,0,1,0,0)
PA<-c(1465,5112,10968,8867,14284)
CA<-c(12000,0,38323,30000,94206)
CH<-c(1,0,1,1,1)
RN<-c(1,2,1,2,2)
VM<-c(7,8,8,8,8)
PoV<-c(2,11,2,11,8,11)
AoV<-c(2,2,1,2,2)
CC<-c(9,7,9,9,7)
df<-data.frame(PN,PC,PA,CA,CH,RN,VM,PoV,AoV,CC)
df
pred2<-predict(LRGmodel,newdata = df)
pred2

```

Annex 3

Script for Naïve Bayes prototype

```
# Two vehicles
```

```
PN<-c("AIC/GTR/MTPV/136676/15","AIC/ADD/MTCM/023499/13")
```

```
PC<-c(1,0)
```

```
PA<-c(1465,5112)
```

```
CA<-c(12000,0)
```

```
CH<-c(1,0)
```

```
RN<-c(1,2)
```

```
VM<-c(7,8)
```

```
PoV<-c(2,11)
```

```
AoV<-c(2,2)
```

```
CC<-c(9,7)
```

```
df<-data.frame(PN,PC,PA,CA,CH,RN,VM,PoV,AoV,CC)
```

```
str(df)
```

```
df
```

```
prd10<-predict(NBmodel,newdata = df)
```

```
prd10
```

Annex 4

List of Risk Items with their numeric representation

Risk Name	Represented by
Ambulance	1
Automobiles-unfamiliarbrand	2
Car-HireBuses	3
Car-HireSmallbuses	4
CarHire-SmallMercides	5
CarHireSmallMiniBUS	6
CarHire-StationWagon	7
CommercialLearnersBus	8
CommercialLearners-MiniBus	9
CommercialLearner-TruckswithTrailer	10
GeneralCartage-DaewooTruck	11
GeneralCartage-FuelTankersTrucks	12
GeneralCartage-FuelTankersTrucksWithTrailers	13
GeneralCartage-IsuzuTrucks-FSR	14
GeneralCartage-IsuzuTrucks-NPR	15
GeneralCartage-Pick-upsOver2350cc	16
GeneralCartage-Pick-upsUpto2350cc	17
GeneralCartage-SINOTipperTruck	18
GeneralCartage-SINOTruck	19
GeneralCartage-TipperFamiliarmodel	20
GeneralCartage-TipperMercedes	21
GeneralCartage-TipperMitsubishi	22
GeneralCartage-TipperNissian	23
GeneralCartage-TruckMercedes	24
GeneralCartage-TruckNissian	25

GeneralCartage-TruckOver100qts	26
GeneralCartage-Trucks	27
GeneralCartage-Truck'sTrailer	28
GeneralCartage-TrucksWithTrailers	29
GeneralCartage-TruckUpto100qts	30
GeneralCartage-WaterTanker	31
IsuzuTruckFSR	32
IsuzuTruckNPR	33
LandCruisers	34
Mercedes	35
Mercedes,BMW&SimilarAuthomobiles	36
Minibuses	37
OwnGoodsTrucks	38
OwnServiceBusesover25seats	39
OwnServiceBusesupto25seats	40
OwnUseCombineHarvestors	41
OwnUseTractors	42
Pickupover2350cc	43
Pickupupto2350cc	44
PrivateCars-Familiarmodel	45
PublicServiceBuses25to45seats	46
PublicServiceBusesupto25seats	47
PublicTransportBusesover45seats	48
PublicTransportIsuzuBuses	49
PublicTransport-MiniBus	50
Truckupto100q	51
TruckwithTrailers	52
VitzYaris	53
WaterTanker	54