



St. Mary's University

Faculty of Informatics

Department of Computer Science

**Loan Risk Prediction Using Machine Learning Algorithms: -The Case
of Ethiopia's Micro-Finance Institution's**

By

Daniel Bizuwork Hailu

Advisor: Dr. Getahun Semeon

June 2019

St. Mary's University
Faculty of Informatics
Department of Computer Science

A Thesis

Submitted to the Faculty of Informatics of
St. Mary's University

In

Partial Fulfillment of the Requirements for the Degree of Master of Science in
Computer science

Loan Risk Prediction Using Machine Learning Algorithms: -The Case of
Ethiopia's Micro-Finance Institute's

By

Daniel Bizuwork Hailu

Advisor: Dr. Getahun Semeon

June 2019

St. Mary's University
Faculty of Informatics
Department of Computer Science

Loan Risk Prediction Using Machine Learning Algorithms: -The Case of
Ethiopia's Micro-Finance Institute's

By

Daniel Bizuwork Hailu

APPROVAL BY BOARD OF EXAMINERS

Chairman Department of Graduate

Signature

Advisor

Signature

Examiner

Signature

External Examiner

Signature

Declaration

I, Daniel Bizuwork Hailu, hereby confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that it has been indicated in the thesis.

Acknowledgment

First and foremost, I would like to give the almighty God who provided me everything to finish this thesis.

I am deeply grateful for my family valued support throughout this Master's program from start to end.

I sincerely thank my thesis advisor Dr. Getahun W/Mariam for his guidance, encouragement and kind personality.

I also would like to thank Aggar, vision fund, Harbu, Pease, Nisir, Oromia and wossasa microfinance institutions and its staffs for their full cooperation for the provision of the required data for my study.

Table of Contents

LIST OF TABLES	i
LIST OF FIGURES	ii
LIST OF ACRONYMES	v
ABSTRACT	vi
CHAPTER ONE.....	1
INTRODUCTION.....	1
1.1 Background	1
1.2. Statement of the Problem	2
1.3 Objectives.....	5
1.3.1. General Objective	5
1.3.2 Specific Objectives	5
1.4. Implementation Tool.....	5
1.5. Significance of the study	6
1.6. Scope and limitation of the study	6
1.6.1 Scope of the study	6
1.7. Organization of the thesis	7
CHAPTER TWO.....	8
REVIEW OF LITERATURE AND RELATED WORKS.....	8
2.1. Introduction	8
2.2. Introduction to Microfinance Institutions.....	8
2.2.1 Micro-Finance Institute in Ethiopia.....	8
2.2.2. Loan Risk in Microfinance Institution.....	9
2.2.3 Probability of Default in MFI.....	9
2.2.4 Financial term.....	9
2.2.5 Loan risk Analysis in MFI.....	10
2.2.6 Loan Disbursement in MFI.....	11
2.3 Machine Learning Techniques	11
2.3.1. Support Vector Machine (SVM)	13
2.3.2. Naive Bayes.....	15
2.3.3. K-Nearest Neighbor (KNN)	16
2.3.4. Logistic regression.....	17

2.3.5. Ensemble Learning.....	18
2.4 Data Mining Research Methodology.....	19
2.4.1 CRISP and SEMA Data Mining Methodology.....	19
2.5. Related work.....	22
CHAPTER THREE.....	36
RESEARCH METHOD AND TECHNIQUES	36
3.1 Introduction	36
3.2 The Cross-Industry Standard Process for Data Mining (CRISP-DM)	36
3.2.1 Business Understanding	37
3.2.2 Data Understanding	38
3.2.3 Data Preparation	38
3.2.4 Predictive Modelling	39
3.2.4.1 Reason for The Chosen Algorithms.....	39
3.2.5. Model Evaluation	40
3.2.5.1 Confusion Matrix.....	41
3.3 Privacy and Confidentiality of Borrowers Data	42
3.4 Summary	43
CHAPTER FOUR	43
DATA PREPARATION	43
4.1 Data Preprocessing	43
4.1.1 Data collection.....	43
4.1.2 Data Cleaning	44
4.1.3 Method of Data Quality Assurance	44
4.1.4 Imbalance Data and Splitting the Data Set.....	44
4.2 Data Transformation.....	45
4.3 Attribute Selection.....	46
4.4 Description of the Data Set.....	47
4.5 Data Transformation.....	48
CHAPTER FIVE.....	51
EXPERIMENT AND DISCUSSION OF RESULTS	51
5.1. Introduction	51
5.2 Data Visualization	52
5.3 KNN Modeling.....	55

5.3.1 Experiment One.....	56
5.3.2 Experiment Two	62
5.3.3 Experiment Three	63
5.3.4 Experiment Four.....	65
5.3.5 Attribute(variable) Importance	66
5.4. Logistic Regression Modeling.....	67
5.4.1. Experiment one.....	67
5.4.2. Experiment two	77
5.5. Navies Bayes Model.....	78
5.6. Support Vector Machine (SVM) Model.....	83
5.6.6 Variable Importance in SVM Model	87
5.7 Summary of Main Findings.....	88
5.7.1 Results on KNN (Nearest neighbor algorithm)	88
5.7.2 Results on SVM (Support Vector Machine).....	92
5.7.3 Results on Logistic Regression Model	95
5.7.4 Results on Naive Bayes Model.....	96
5.7.5 Machine Learning Models Comparison	97
5.8. Summary	98
CHAPTER SIX	99
CONCLUSION AND RECOMMENDATION	99
6.1 Conclusion.....	99
6.2 Recommendations and Future Works.....	100
REFERENCES	102
Appendix I.....	109
Appendix II.....	111

LIST OF TABLES

Table 2.1: - Summary of literatures reviewed on loan risk prediction.....	30
Table 3. 2: Example of Confusion Matrix.....	41
Table 4. 1: Attributes of data set used for MFIs loan risk prediction.....	47
Table 4.2: Variables and transformed data types.....	49
Table 5.1 Summary of KNN experiments result.....	89
Table 5.2 Confusion matrix result of KNN experiment one.....	89
Table 5.3 confusion matrix of KNN experiment two.....	89
Table 5.4 Confusion matrix of SVM experiment.....	93
Table 5.5 Confusion matrix of logistic regression experiment.....	95
Table 5.6 Confusion matrix of logistic regression experiment.....	96
Table 5.7 Confusion matrix of Naive Bayes experiment.....	97
Table 5.8 Confusion matrix of Naive Bayes experiment.....	97

LIST OF FIGURES

Fig. 2.1 Machine learning tasks	13
Fig. 2.2 A linear line separating the data types.....	14
Fig 2.3 General Logit Curve.....	18
Fig 2:4 The CRISP-DM process model.....	20
Fig 3:2 Confusion Matrix.....	39
Fig 4.1 Transformed dataset.....	51
Fig 5.1 Frequency of distribution of Age.....	52
Fig 5.2 Frequency of Approved Loan Amount.....	53
Fig 5.3 Frequency of Percentage of Collateral Ownership.....	53
Fig 5.4 Frequency of Loan Cycle.....	53
Fig 5.5 Frequency of Loan term in month.....	54
Fig. 5.6: Plot of loan status vs loan cycle.....	54
Fig. 5.7: Plot of loan status vs loan term.....	55
Fig. 5.8: Plot of loan status vs loan product.....	55
Fig. 5.9: The data set snap shot.....	56
Fig. 5.11: The data structure and summary of the data set snap shot.....	58
Fig. 5.12: The cross table result snap shot of experiment one.....	60
Fig. 5.13: The cross table result snap shot of experiment two.....	62
Fig. 5.14: The cross table result snap shot of experiment three	64

Fig. 5.15: The training model of different K values and result snap shot.....	65
Fig. 5.16: The training model of different K values and result plot.....	66
Fig. 5.17: Attribute importance snap shot.....	66
Fig. 5.18: Attribute importance plot.....	67
Fig. 5.19: The data set view snap shot.....	68
Fig. 5.20: The structure of the data set view snap shot.....	68
Fig. 5.21: The result of logistic regression model snap shot.....	71
Fig. 5.22: The result of logistic regression model with significant attributes snap shot.....	73
Fig. 5.23: Cross table result snap shot.....	74
Fig. 5.24: Confusion matrix results snap shot.....	75
Fig. 5.25: Confusion matrix results snap shot.....	76
Fig. 5.26: ROC Curve result snap shot.....	76
Fig. 5.27: Cross table and confusion matrix results snap shot.....	78
Fig. 5.28: The data set and its structure snap shot.....	79
Fig. 5.29: The cross table result snap shot.....	81
Fig. 5.30: Confusion matrix results snap shot.....	82
Fig. 5.31: Structure of the data set snap shot.....	83
Fig. 5.32: Confusion matrix snap shot.....	86
Fig. 5.33: Variable importance ROC Curve in SVM model snap shot.....	87
Fig. 5.34: Variable importance plot in SVM model snap shot.....	87

Fig 5. 35: Attributes contribution to the KNN prediction model.....90

Fig 5.36: Total prospective borrowers in MFIs type.....91

Fig 5.37: Active borrowers in MFIs type.....91

Fig 5.38: Defaulter borrowers in MFIs type.....91

Fig 5.39: data on business sector.....92

Fig 5.40: Defaulter borrowers based on sector.....92

Fig 5.41: active borrowers based on sector.....92

Fig 5.42: Attributes contribution to the SVM prediction model.....93

LIST OF ACRONYMES

MFI	Micro-Finance Institute
NBE	National bank of Ethiopia
KNN	K-nearest neighbor
SVM	Support vector machine
PAR	Portfolio's at Risk
ML	Machine learning
NB	Naïve Bayes
NGO	Non-Governmental Organization
LG	Logistic Regression
MFILD	Microfinance Loan Data
CEO	Chief Executive Officer
AIC	Akaike information criterion
CSV	Comma-separated values

ABSTRACT

Due to the wide availability of computer, information and communication technologies data are being generated massively today, especially in financial institutions and banks data are being generated massively on regular basis. Microfinances are one of such institutions that collect, process and store huge amounts of records from time to time and therefore deal with large amount of data. On the other hand, Ethiopian Microfinances are facing problems in loan risk assessment and managing portfolio at risk. Currently Ethiopian microfinance institutions loan risk assessment and granting loan to the borrower's is conducted in a traditional manner depends on the loan approval team views and believes, Moreover, such way of risk assessment creating inefficiency in quality of identifying borrower's characteristics before granting the loan. If the microfinance institutions (MFIs) do not manage their loan risks well, they are likely to fail to meet their social and financial objectives. The existing past and historic data related to loan borrower and loan characteristics could be actionable and usable for loan risk assessment with the help of Machine learning algorithms. This study was conducted to demonstrate the practical methods, experiments and datasets with machine learning to assist Ethiopian MFIs through building a classification and prediction model which supports in prediction of a new loan borrower's status (Active or Defaulter) when the loan decision making in the microfinance institutions. The classification and prediction model are built based on the MFIs loan borrowers' data obtained from the selected seven (Aggar, Harbu, Vision Fund, Pease, Oromia, Nisir and wosasa) microfinance institutions in Ethiopia. Necessary preprocessing activities have been applied to clean and make it ready for the Experimentation. Then, the four algorithms used were SVM, KNN, Naïve Bayes and logistic regression. The RStudio with R programming was used to simulate all the experiments. Confusion matrix was used to calculate the accuracy, specificity, sensitivity and precision were used to evaluate the performance of the models and Cross table was used to visualize the performance of the models. The results of the experiment show high precision, so that the models can be used in detecting and predicting defaulter (risky) loan applicants. The KNN classifier produced an accuracy of 99.91%, the SVM classifier produced an accuracy of 92.4%, logistic regression model also produced an accuracy of 93.8%, and Naïve Bayes classifier produced an accuracy of 83.8 %.

Keywords: Machine learning algorithms, loan risk assessment, MFIs

CHAPTER ONE

INTRODUCTION

1.1 Background

Due to the wide availability and increasing processing power and the continuous decline in the cost of storage devices, data are being generated massively today than it was decades ago [8]. With fields in the financial industry like bank and microfinance institutes, data are being generated massively on regular basis. So, such financial institutions are finding ways to turn these data into very beneficial information to their advantage [8]. Loan risk is considered the most serious problems of every MFIs [12]. Past events contain patterns that are hidden in the data that records them. Machine learning promises the use of models to go through and analyze huge data that are very difficult for human experts, and also possesses incredible power to detect hidden relationships, correlations and associations in data [23]. This study conducted experiments using machine learning algorithms to predict the outcome of loan risk using MFIs datasets.

Micro-Finance Institutions (MFIs) are categorized as a financial service institution to provide finance for individuals and small businesses. The two main way of loan in microfinance are loans to groups whose members use social capital to screen out risks, and loans to individuals whose loan officers know them well enough to screen out bad risks, rely fundamentally on qualitative information held in human memory [7]. Loan Risk prediction and classification, in contrast, relies fundamentally on quantitative information stored in micro finance institutes database or excel format in a computer system.

Loans default will cause huge loss for the financial institution's, so they pay much attention on this issue and apply various method like loan default prediction to detect and predict default behaviors of their customers [9].

If the microfinance institutions (MFIs) do not manage their loan risks well, they are likely to fail to meet their social and financial objectives [43]. As with any financial institution, the biggest risk in microfinance is a loan default risk [12].

Loan risk prediction is used to predict the probability of a given borrower defaulting on his or her loan. The default probability is predicted based on the influence of defined risk factors on loan repayment performance of a cluster of similar clients in the past.

1.2. Statement of the Problem

Managing financial institutions specially microfinance has never been easy, but in recent years it has become even more difficult because of greater uncertainty in the economic environment [43]. Loan risk management is one of the most important activities in any financial institution's and cannot be overlooked by any economic enterprise engaged in loan irrespective of its business nature. If the microfinance institutions (MFIs) do not manage their loan risks well, they are likely to fail to meet their social and financial objectives. When poorly managed risks begin to result in financial losses, donors, investors, lenders, borrowers and savers tend to lose confidence in the organization and funds begin to dry up. When funds dry up, microfinance institutions (MFIs) are not able to meet their social objective of providing services to the poor and quickly go out of business [43]. As with any financial institution, the biggest risk in microfinance is default risk [12].

Microfinance Institutions (MFIs) provide microcredits, small loans, to low income individuals. Like every loan, they must be reimbursed. For this reason, the MFI must assess and evaluate the financial aspects as well as the risks of the operation [11]. The aim of this loan risk prediction and classification is to assess the creditworthiness of the applicant, that is the main obstacle for MFIs to provide loans to clients. This finally result loan risk which is the one that negatively affect the performance of MFIs [12].

In Most MFIs that specialize in individual lending, on average loan officers spend 40% to 50% of their time in collection activities [13]. Whereas loan officers in MFI face challenges in control and collecting overdue payment. As a result, they need other operation strategies such as Loan risk prediction and classification Model that can reduce time spent in collecting overdue payments from delinquent borrowers and also can help reduce the time by prioritizing the visits to those borrowers who are more likely to default, leaving loan officers more time to identify and access new customers or promoting products to the customers [13].

In a microcredit application, financial information is scarce because the applicants do not maintain bookkeeping, and they generally lack of credit history [11]. Microfinance lenders, however, do not have access to credit bureaus, and most of their borrowers are small business owner and self-employed [14].

In Ethiopian National Bank, there are credit reference bureaus responsible to bank supervision that use a system called credit registry system to register loan borrowers in all Ethiopian commercial banks, that inform a client has a defaulter or not, based on past loan history. However, this registry system is not available and applicable for microfinance institutes. Loan Applicants in MFI do not generally have loan history records but currently more than 5.3 million borrowers are active. Furthermore, MFIs face extra challenges, the loan applicant risk assessment is done without such applicant loan history. In order to win this challenge, it is required to analyses the risk and predict the probability of default for a given loan applicant based on the influence of defined risk factors on loan repayment performance in a cluster of similar clients in the past before loan disbursement.

The sustainability of microfinance institutions depends largely on their ability to collect their loans as efficiently and effectively as possible [12].

In other words, to be financially viable or sustainable, microfinance institutions must ensure high portfolio quality based on total repayment, or at worst low delinquency/default, cost recovery and efficient lending [25].

From the interviews and discussions made with senior managers of the MFI institutions the existing loan risk analysis and loan granted is dependent on loan officer views of borrowers and loan applicant commitment form of borrowers, it requires significant improvement to minimize the loan risk and identify defaulter borrower's in early stage before loan disbursement.

Almost all selected microfinance institutions are at risk regarding high rate of default/delinquency by their clients; which are most of microfinance institutions are not achieving the internationally accepted standard portfolio at risk of 5% [25], which is a cause

for concern because of its consequences on MFIs businesses, individuals, and the economy of Ethiopia at large.

The effective way to do this and meet financial sustainability by applying Machine learning which makes another possibility through designing a learning, classification and prediction phases. Machine learning plays a major role in computer science research, and such research has already had an impact on real-world applications [23]. It assists Loan Approvals decision makers in MFIs as loan risk prediction and classification tools and also it allows MFIs to manage their portfolio at risk ratios more precisely.

Other studies conducted in MFIs like [46] [37], the attributes used for their study is very limited in number that is 7 and 10 respectively and only about one loan product characteristics and they exclude the loan borrower's characteristics for their consideration. With one microfinance institutions, the data set also very limited that is 4000 and 6447 records and study is not inclusive to other microfinance institutes.

In this study the problem of loan risk prediction was addressed using borrower's characteristics which are significant in real world loan risk assessment practice but not explored more in previous studies and loan characteristics. As stated in [50], loan characteristics, business characteristics and loan borrowers' characteristics attributes used in Ethiopian MFIs for loan risk assessment (to identify good and bad borrower's) like sex, age, monthly income earns, occupation, business type, location of collateral, business location of borrower's, and education status of borrowers are included in this study.

This research therefore aims to develop loan risk prediction model using machine learning algorithms for MFIs in Ethiopia. The research focus on collecting and analysis of loan application data to accurate risk assessment and to manage portfolio at risk ratios precisely.

To this end, this study attempts to explore and answer the following question: -

1. What is the extent of loan risk for MFI in Ethiopia?
2. How best can the captured knowledge be utilized in making decision whether to extend loan or not?
 - 2.1 Which attributes are most useful in predicting loan risk?

2.2 We selected four machine learning models namely: SVM, Navies Bayes, K-NN and logistic regression. Which model is more accurate for predicting loan risk?

1.3 Objectives

The following are general and specific objectives of the study.

1.3.1. General Objective

The general objective of the study is to analyze existing loan risk assessment and predict loan risk for accepting or rejecting loan application using machine learning algorithms.

1.3.2 Specific Objectives

- To prepare appropriate data sets with relevant attributes for classification and prediction model.
- To identify and find the main attributes that can help to predict loan risk.
- To assess the usability of different algorithm and techniques used so far in other researches for the prediction of loan risk.
- To identify models, techniques and implementing tools relevant for the loan risk prediction.
- To conduct experiments on loan risk prediction system for identifying loan defaulter using a specific implementation tool
- To evaluate loan risk prediction model using performance measures on the bases of its accuracy and precision.
- To forward recommendations for further research.

1.4. Implementation Tool

For this study, the experimentation process was done using R studio with R programming. R is an integrated suite of software facilities for data manipulation, calculation and graphical facilities for data analysis and display, it preferred because of the following reasons: R-Studio has easy-to-use Interface-Studio has ability to present datasets in the form of figures with variety of presentations, in R-Studio, it is possible to link datasets in common simple database format such as .CSV and R programming is simple and suitable for technical computing [71].and also R is with Effective data handling and storage ,Suite of operators for calculations

on arrays and Large, coherent, integrated collection of intermediate tools for data analysis • Programming language, run time environment [71].

1.5. Significance of the study

This study makes several contributions to both knowledge building and practice improvement in loan risk prediction. From this study, Micro-Finance Institute's in Ethiopia are mainly beneficiaries: -

- For identifying and detect possible defaulter and active loan applicants before granting loan to the borrowers.
- To limit or to completely end from granting loans to risky loan.
- To instantaneously approving low-risk customers and secure PAR (portfolios at risk) as internationally accepted level below 5%.

In addition to this, the study contributes to knowledge building including for other researchers interested in similar area.

- The study identifies different and massively contributes attributes in microfinance institutions loan risk assessment, those are Percentage of collateral ownership, business sector attributes, Business Yearly Earnings attributes, Education status attributes, and MFIs organization Type attributes.
- The study shows that Machin learning algorithms can be apples to Ethiopian microfinance institutes for loan risk assessment with different and relatively big data set 37,380 records, with more attributes 18 attributes and a better accuracy 99.91% and inclusive to all MFIs model than prior study related to microfinance like [46] [37].
- Plus, the study gives good understanding in the concepts of Machine Learning by building models using R.

1.6. Scope and limitation of the study

1.6.1 Scope of the study

The proposed research topic is aimed to investigate and identify patterns which help to Predict the probability of default for a given loan applicant through model, and Classify to accept or

reject loan application using borrowers historical MFIs data set. This work encompasses and pass through from preprocessing the loan and borrower's historical data to their proper prediction and classification of the loan applicants. The loan and borrower's data collect from selected microfinance in Ethiopia that were serve for learning, prediction and classification purpose.

1.7. Organization of the thesis

The thesis is organized as follows with six chapters. The first chapter is about introductory part which gave an overview of background and statements of the problem for the study, objectives, scope and limitations of the study, significance of the study and description about the methodology to conduct the study.

Chapter two of this study talks about literatures and related works reviewed. All the reviewed literature was in the light of machine learning in the financial sector. The sections are machine learning and machine learning techniques (Classification, Prediction, Clustering, Description and visualization), Importance and application of machine learning, how machine learning can assist in loan risk assessment in the microfinance institutions, type of risks and risk assessment in microfinance is presented.

Chapter three embodies the methodology of the study. The following headings are discussed. Research methodology, the CRISP-DM Process, model evaluation methods (confusion matrix) and privacy and confidentiality of loan borrower's data.

Chapter four presents data preparation, in this chapter the following topics are discussed; data preprocessing, description of the data set and attribute selection.

Chapter five presents experiment and result discussion. It discusses how the four models were developed and evaluated. And also deals with results and analyses of the models.

Chapter six is about the conclusion, recommendations and future works.

CHAPTER TWO

REVIEW OF LITERATURE AND RELATED WORKS

2.1. Introduction

This chapter presents conceptual discussion on loan risk and analysis in microfinance and a comprehensive review of literature related to application of machine learning techniques in loan risk prediction published in academic journals.

2.2. Introduction to Microfinance Institutions

Microfinance refers to the provision of small-scale financial services including microcredit, savings, payment services, micro insurance and other services to the rural and urban clients [1].

In Ethiopia, microfinance services were introduced after the demise of the Derg regime following the policy of economic liberalization [48]. These MFIs aim at financing to micro and small level enterprises, individuals and agricultural business in individual and group-based lending [48].

2.2.1 Micro-Finance Institute in Ethiopia

The delivery of financial services in Ethiopia through the MFI has increased in a short period of time. Both outreach and sustainability of the microfinance institutions have increased significantly. Currently thirty-five microfinance institution are on action under supervision of national bank of Ethiopia NBE Repost 2017/2018.

Ethiopian microfinance sector is characterized by its rapid growth, an aggressive drive to achieve scale, a broad geographic coverage, a dominance of government backed MFIs, an emphasis on rural households, the promotion of both credit and savings products, a strong focus on sustainability and by the fact that the sector is Ethiopian owned and driven [10]. According to micro-financing business proclamation No. 626 /2009 the purpose and activity of MFIs in Ethiopia are to collect deposits and extend credit to rural and urban farmers and people engaged in other similar activities as well as micro and small scale rural and urban entrepreneurs.

2.2.2. Loan Risk in Microfinance Institution

Loan risk is defined as the probability that some of MFI 's assets, especially its loans, will decline in value and possibly become worthless. Because MFIs hold little owners 'capital relative to the aggregate value of their assets, only a small percentage of total loans need to go bad to push MFI to the brink of failure. Thus, management of Loan risk is very important and central to the health of MFI and indeed the entire financial system. As MFI s makes loans, they need to make provisions for loan losses in their books. The higher this provision becomes, relative to the size of total loans, the riskier MFI becomes. An increase in the value of the provision for loan losses relative to total loans is an indication that the MFI 's assets are becoming more difficult to collect [43]. Loan risk is the risk of a loss resulting from the debtor's failure to meet its obligations to the MFI in full when due under the terms agreed [44].

2.2.3 Probability of Default in MFI

Reviewing a borrower's probability of default is basically done by evaluating the borrower's current and future ability to fulfill its interest and principal repayment obligations. A default is an event that a borrower cannot meet his contractual obligations [25]. The exact definition of what this means differs from institution to institution. Basically, the framework, a microfinance loan is considered to be at risk that is three months past due, is considered as a default.

2.2.4 Financial term

Defaults: -This study more deal about defaults and the probability that they occur. A default is an event that a borrower cannot meet his contractual obligations. The exact definition of what this means differs from institution to institution. Basically, the framework, every loan that is three months past due, is considered as a default. The probability that a certain loan defaults in some time horizon (mostly two year or the lifetime of a loan) is called the probability of default [25].

Delinquency: -Payment delinquency is commonly used to describe a situation in which a borrower misses their due date for a single scheduled payment for a form of financing, like student loans, mortgages, credit card balances, or automobile loans. There are

consequences for delinquency, depending on the type of loan, the duration, and the cause of the delinquency [25].

2.2.5 Loan risk Analysis in MFI

Credit analysis is the primary method in reducing the credit risk on a loan request. This includes determining the financial strength of the borrowers, estimating the probability of default and reducing the risk of non-repayment to an acceptable level.

In general, loan evaluations are based on the loan officer's subjective assessment (or judgmental assessment technique). Loan analysis is essentially default risk analysis, in which a loan officer attempts to evaluate a borrower's ability and willingness to repay [43]. Lawrence stated that A MFI's loan risk analysts often use the five C's as key Dimensions of an applicant's loan worthiness [43] [44]. They include; Character, Capacity, Capital, Collateral, and Conditions.

Character: The applicant's record of meeting past obligations, financial, contractual, and moral. Past payment history as well as any pending or resolved legal judgments against the applicant would be used to evaluate its character.

Capacity: The applicant's ability to repay the requested credit. Financial statement analysis, with particular emphasis on liquidity and debt ratios, is typically used to assess the applicant's capacity.

Capital: The financial strength of the applicant as reflected by its ownership position. Analysis of the applicant's debt relative to equity and its profitability ratios are frequently used to assess its capital.

Collateral: The amount of assets the applicant has made available for use in securing the credit. The larger the amount of available assets, the greater the chance that a firm will recover its funds if the applicant defaults. A review of the applicant's balance sheet, asset value appraisals, and any legal claims filed against the applicant's assets can be used to evaluate its collateral.

Conditions: The current economic and business climate as well as any unique circumstances affecting either party to the credit transaction. For example, if the firm has excess inventory of

the items the applicant wishes to purchase on credit, the firm may be willing to sell on more favorable terms or to less creditworthy applicants. Analysis of the general economic and business conditions, as well as special circumstances that may affect the applicant or firm is performed to assess conditions.

2.2.6 Loan Disbursement in MFI

Loan disbursement is the act of giving or paying out money to customers who have been accessed and approved to be given loan [44]. After an applicant has been carefully assessed and has been proven that the applicant meets the loan requirement of the microfinance institutions. The loan officer together with the loan committee gives their approval by appending their signature on the loan application form. This gives bank the right to disburse the funds to the applicant. Disbursement ensures that money is made available to the customer after all assessment has been done and approval has been given. The assessment process also ensures that the authenticity of the security and other required documentations are received certified before funds are given out to the qualified customer [43].

2.3 Machine Learning Techniques

Now a day's large amount of varieties and velocity of data are generated and available everywhere specially in financial sector, therefore it is important to analysis this data in order to extract some useful information and to develop an algorithm based on this analysis.

There are many machine learning algorithms that can be used to classify a problem given a set of features. These work looks to those algorithms to see if any are particularly useful in classifying Ethiopian microfinance institutes data to predict loan defaulters.

At this time more powerful computational tools have opened a new frontier in the field of data science. And thus, there is currently an active ongoing research within the fields of machine learning (building analytical models using algorithms for machine to “learn” from the data) [28].

The most important background information on machine learning algorithms and their theoretical formulation are outlined in this section.

Machine learning is an integral part of artificial intelligence which is used to design algorithm based on the data trends and historical relationship between the data [36].

Machine learning is an emerging technique for building analytic models for machines to “learn” from data and be able to do predictive analysis. The ability of machines to “learn” and do predictive analysis is very important in this era of big data and it has a wide range of application areas. For instance, banks and financial institutions are sometimes faced with the challenge of what risk factors to consider when advancing credit/loans to customers [28].

Machine learning techniques can be grouped broadly into two main categories. They include: Supervised Learning and Unsupervised Learning [28].

The main feature of supervised learning algorithm consists of target or outcome variable (or dependent variable). The target variable is used to predict other features from a given set of predictors (independent variables). Furthermore, using the target variable, a function is generated that maps input to desired outputs. The training process then continues until the model achieves the desired level of accuracy on the training data. Such supervised learning techniques are achieved using regression and classification algorithms or approaches that range from non-linear regression, generalized linear regression, discriminant analysis, naive Bayes, logistic regression, Support Vector Machines (SVMs) to decision trees and k-nearest neighbor algorithm [28].

In unsupervised learning, there is no target or outcome variable to predict or estimate. This algorithm is used mainly for segmenting or clustering entities in different groups for specific intervention. Examples of unsupervised learning algorithms include Neural Networks and K-means algorithms [28].

The following figure shows various machine learning approaches and algorithms.

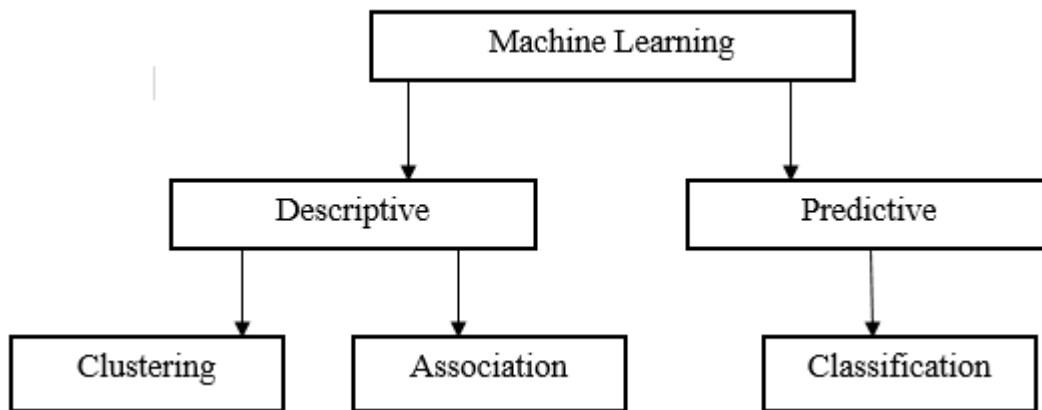


Fig. 2.1. Machine learning tasks.

There are many machine learning algorithms that can be used to classify a problem given a set of features in both supervised and unsupervised learning. These work looks to supervised machine learning algorithms to see if any are particularly useful in classifying microfinance institutions loan transactional data and give probability of default prediction. Algorithms investigated in this study are: Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), naive Bayes and logistic regression.

2.3.1. Support Vector Machine (SVM)

Support Vector Machines are supervised learning models that can be used for classification, prediction and clustering problems. An SVM takes a set of input observations and associated binary outputs and constructs a model that can classify new observations into one class or the other. Support vector machine is highly preferred by many as it produces significant accuracy with less computation power [29].

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side [30].

The key element is that support vector machine algorithm is to find a hyperplane in an N-dimensional space (N—the number of features) that distinctly classifies the data points.

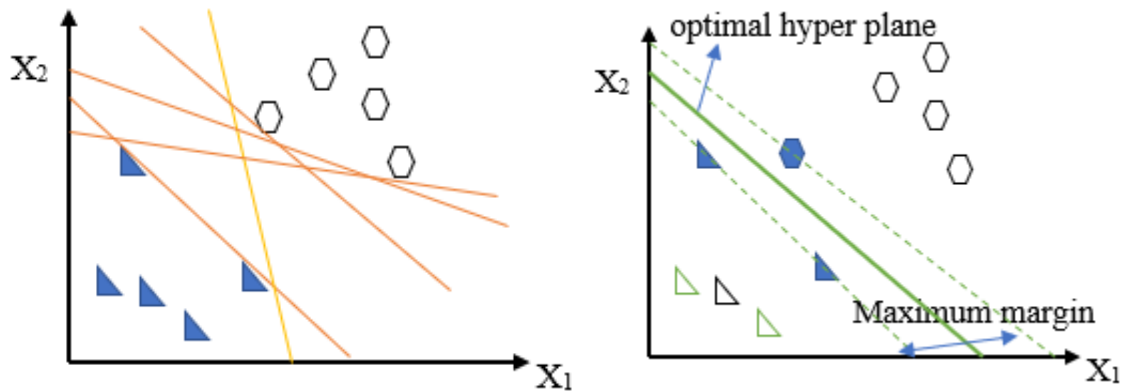


Fig. 2.2 A linear line separating the data types

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The main thing in SVM is to find a plane that has the maximum margin, that means the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

In real world application, there is trade off on finding perfect class when the two classes are not linearly separable like due to noise, the condition for the optimal hyper-plane can be relaxed by including an extra term: millions of training data set.

In the case of real-world application, it is not usually possible to get a line that perfectly separates the data within the space. Hence, we might have to use a curved decision boundary. It is possible to get a hyper-plane which could separate the data but this may not be desirable if the data has noise in it. In such cases we need to use the soft margin method [30]. The soft margin method allows for points to appear on the incorrect side of the margin. These points have a penalty associated with them. The penalty increases as the points are farther from the margin. The hyperplane separation looks to minimize the penalty of incorrectly labeled points, while maximizing the distance between the remaining examples and the margin.

The other approach is SVM kernel it employed to separate data that isn't linearly separable, is to map the data into a higher dimensional space. By mapping $x = (x, x_2)$ the data will be mapped into two-dimensional space. When this two-dimensional mapping is graphed, an obvious

linearly separable line appears [30]. The mapping used to increase the dimensionality of the problem is dependent on the data space being investigated. The above computations, which are used to find the maximum-margin separator, can be expressed in terms of scalar products between pairs of data points in the high-dimensional feature space. These scalar products are the only part of the computation that depends on the dimensionality of the high-dimensional space.

2.3.2. Naive Bayes

Naive Bayes is a widely used classification method based on Bayes theory. Based on class conditional density estimation and class prior probability, the posterior class probability of a test data point can be derived and the test data will be assigned to the class with the maximum posterior class probability [34]. Naive Bayes algorithm is one of the most effective methods and is a simple but surprisingly powerful algorithm for predictive modeling [33].

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. [34]

In Naive Bayes probability theory, Bayes theorem relates the conditional and marginal probabilities of two random events [34]. It is often used to compute posterior probabilities given observations. Let $x = (x_1, x_2, \dots, x_d)$ be a d -dimensional instance which has no class label, and our goal could be to build a classifier to predict its unknown class label based on Bayes theorem. Let $C = \{C_1, C_2, \dots, C_K\}$ be the set of the class labels. $P(C_k)$ is the prior probability of C_k ($k = 1, 2, \dots, K$) that are inferred before new evidence, $P(x|C_k)$ be the conditional probability of seeing the evidence x if the hypothesis C_k is true. A technique for constructing such classifiers to employ Bayes' theorem to obtain is given by the following formula: -

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{k'} P(x|C_{k'})P(C_{k'})} \quad [1]$$

A naive Bayes classifier assumes that the value of a particular feature of a class is unrelated to the value of any other feature, so that [34]:

$$P(x|C_k) = \prod_{j=1}^d P(x^j|C_k) \quad [2]$$

2.3.3. K-Nearest Neighbor (KNN)

The nearest neighbor (NN) classifiers, especially the k-NN algorithm, are among the simplest and yet most efficient classification rules and are widely used in practice [30].

The purpose of KNN algorithm is to use a database in which the data points are separated into several separable classes to predict the classification of a new sample point. An object is classified by looking to its nearest examples [31]. In KNN algorithms, K states how many neighbors will be used in voting, it is important to decide the value of k because the accuracy of the classification is dependent on it. K=1 simply states that an object will be assigned the same class as its nearest example. As the number of K increases then we need to classify a given instant based on the resemblance of all the stated K instances.

The measurement can be performed using any distance metric or similarity function, such as the Euclidean, Cosine or Jaccard [31]. The classifier is developed from the training data documents in this case and to check for the efficiency and accuracy of the classifier the holdout method has been used. In the holdout method the original training dataset is partitioned into two, a major portion to generate the classifier and the minor portion to check the precision of the classifier. It is an assumption that documents of the same class will always be the nearest neighbors of each other i.e. the distance between them will be less as they are in some way related to each other [31].

The major problem of using the k-NN decision rule is the computational complexity caused by the large number of distance computations and the other most critical problem is Selecting the value of K in K-nearest neighbor. A small value of K means that noise will have a higher influence on the result i.e., the probability of overfitting is very high. A large value of K makes it computationally expensive and defeats the basic idea behind KNN (that points that are near might have similar classes). A simple approach to select k is $k = n^{1/2}$ [32].

In the general process of KNN algorithms the first step is preprocessing the data in the database in such a way as to ensure that we can compare observations. Then, our observations become points in space and we can interpret the distance between them as their similarity (using some appropriate metric) One of the most widely used metrics is the Euclidean distance. The Euclidian distance between two instances $(X_1, X_2, X_3, \dots, X_n)$ and $(U_1, U_2, U_3, \dots, U_n)$ is given by the following formula: -

$$\sqrt{(X_1 - U_1)^2 + (X_2 - U_2)^2 + \dots + (X_n - U_n)^2} \quad [3]$$

When we take loan rating as an example using KNN algorithms, first we collect financial characteristics and comparing people with similar financial features to a database. By the very nature of a loan rating, people who have similar financial details would be given similar loan ratings. Therefore, we will like to be able to use this existing database to predict a new customer's loan granted, without having to perform all the calculations.

2.3.4. Logistic regression

Logistic regression is also called logistic model or logit regression. It is a predictive analysis. It takes independent features and returns output as categorical output. The probability of occurrence of a categorical output can also be found by logistic regression model by fitting the features in the logistic curve [35].

Logistic Regression, falls under Supervised Machine Learning. It solves mainly the problems of Classification to make predictions or take decisions based on past data [35]. It is used to predict binary outcomes for a given set of independent variables. The dependent variable's outcome is discrete.

The output of Logistic Regression is a sigmoid curve or more popularly known as S-curve. Where the value on the x-axis, independent variable would determine the dependent variable on the y-axis. In logistic regression there are only two possible outcomes. 0 and 1. That something occurs, or it doesn't. We use a threshold value to make our prediction easier. If the x-axis' corresponding y-value probability is lesser than the threshold value, the outcome is taken as 0. If it is greater than the value, the outcome is taken as 1.

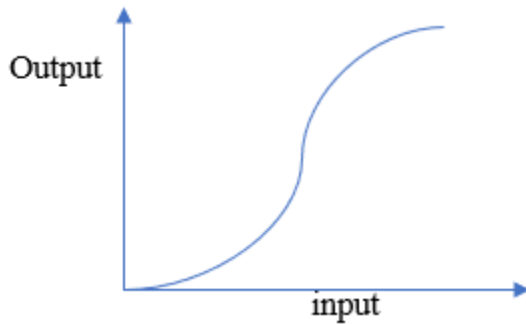


Fig 2.3 General Logit Curve

The Logistic Regression model can be replaced by the simpler Linear Regression model when the output variable is taken to be continuous. When the output variable is not continuous or is dichotomous another model has to be applied in order to take this difference into consideration.

Logistic Regression model was chosen over the other models because of its mathematical clarity and flexibility. This model can have single or multiple predictors [35].

2.3.5. Ensemble Learning

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their classification [41]. Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem [45]. The original ensemble method is Bayesian averaging but more recent algorithms include error-correcting output coding, bagging and boosting. Because uncorrelated errors of individual classifiers can be eliminated through averaging ensembles are often much more accurate than the individual classifiers that make them up. There are different methods of constructing ensembles namely [41]; Manipulating the Training Examples, Manipulating the Input Features, Manipulating the Output Targets and Injecting Randomness.

The first method manipulates the training example to generate multiple hypotheses [41]. The learning algorithm is run several times each time with a different subset of the training examples. This technique works especially well for unstable learning algorithms whose output classifier undergoes major changes in response to small changes in the training data. The

second method manipulates the set of input features available to the learning algorithm [30]. The method usually only works when the input features are highly redundant. The third method manipulates the y values that are given to the learning algorithm [46]. Having a large class, K, new learning problems can be constructed by randomly portioning the K classes into two subsets A and B and give a level of 0 and 1 respectively for the subsets. The relabeled data from the subsets then given to learning algorithm which then can construct the classifier. The last method injects randomness into the learning algorithm [41].

2.4 Data Mining Research Methodology

2.4.1 CRISP and SEMA Data Mining Methodology

2.4.1.1 The CRISP-DM Process

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. As a process model, the CRISP-DM gives a general outline of the data mining life cycle. There are six phases in the CRISP-DM data mining life cycle. There are arrows that link the various phases with no exact order. The arrows show that, the various phases of the DM life cycle are interdependent. This means that the arrows depict reliance of the phases. For some projects to be successful there is the need to be switching between two or more phases before completion, others may not necessarily need that. The outermost circle shows data mining itself is revolving [69]. The following figure 2.4 shows the six phases of the CRISP-DM process model and their interactions.

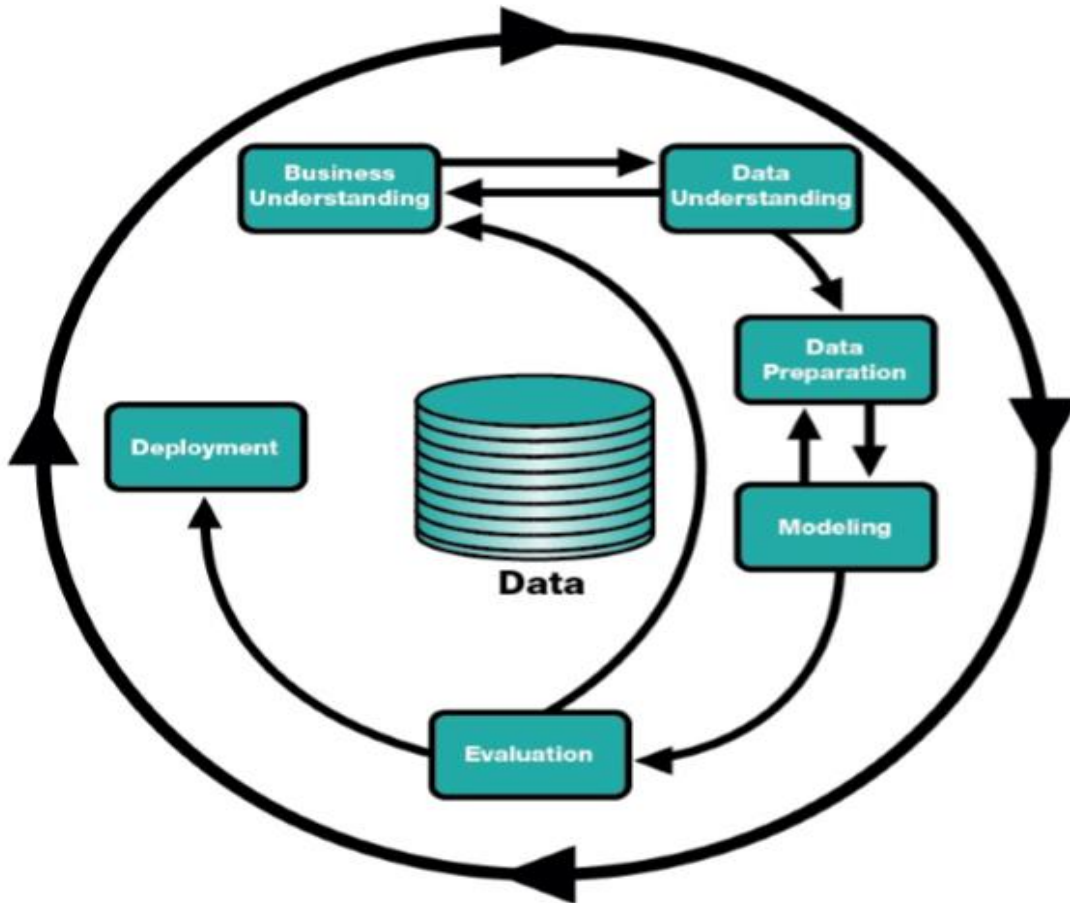


Figure 2.4: The CRISP-DM Process, figure source [69].

Business understanding-this first phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.

Data understanding-the data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Data preparation-the data preparation phase covers all activities to construct the final dataset from the initial raw data.

Modeling-in this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

Evaluation-at this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives.

Deployment-creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it [72].

2.4.1.2 The SEMMA Process

SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a DM project. It is data mining method developed by SAS Institute considers a cycle with five stages for the process, those are:

Sample - this is the first stage of the SEMMA process, consists on sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly.

Explore - this is the second stage of the SEMMA process, consists on the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify - this is the third stage of the SEMMA process, consists on the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

Model - this is the fourth stage of the SEMMA process, consists on modeling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Assess – this is the fifth stage of the SEMMA process, consists on assessing the data by evaluating the usefulness and reliability of the findings from the DM process and estimate how well it performs.

The SEMMA process offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for his conception, creation and evolution, helping to present solutions to business problems as well as to find de DM business goals [72].

2.5. Related work

Researches have been working on loan risk assessment and loan risk prediction using the advantage of Machine Learning models and data mining. Different related researches have been conducted using different data mining techniques, machine learning algorithms, and tools on various loan predicting loan default and loan risk assessment in financial sectors locally and internationally. The following is an overview of the papers that served as guides to this work. The overview covers the four selected models (SVM, Navies Bayes, K-NN and logistic regression) and in addition it looks for ensemble models and other algorithms published in recent academic journals.

Paulius Danenasa, and Saulius Gudasc [40], conducted an empirical evaluation of SVM-based classifiers applied for credit risk evaluation task in the real-world dataset. It is argued that the methodology presented in this study could serve as an alternative tool for risk analysis in case when there are no actual bankruptcy classes or obtaining them might be a too complicated or expensive. The dataset was evaluated in three ways: as original (unchanged) dataset 3266 entries with 13 number of attributes and transformed into differences expressed as absolute 1912 with 11 number of attributes and percentage values 1912 with 10 number of attributes respectively.

Experiment results showed that feature selection resulted even in better accuracy with which bankruptcy class can be predicted with accuracy ranging from 82 to 84%; however, the study shows efficiency of SVM classifiers much depends on parameter selection, thus obtained results can be improved by proper parameter selection, another aspect especially important in this model application is dataset balance as classes are computed dynamically.

Researcher stated that SVM is one of the machine learning techniques which is sensitive to dataset imbalance as “majority” classes tend to outweigh “minority” classes by pushing classification boundaries over them.

This research concludes that SVM classifiers, together with gradient descend based SVM classifiers and Core Vector Machines algorithms, can be a good choice or alternative for implementation of SVM-based credit risk evaluation model.

The study conducted by Vimala S. and Sharmili K [39], focused on a title - Prediction of loan risk model by Combination of Naïve Bayes and Support Vector Machine on using data from banking sector. To conduct the experiment, they used dataset from UCI (Machine learning data set repository), and provided an original dataset germen credit (credit service company dataset repository). This dataset has 850 records with 21 attributes with their respected class. They used a combination of Naïve Bayes and Support vector machine for the prediction. The accuracy of the model (combination NBSVM) stated by the author is 81% in average experiments.

From these classification techniques the study shows that Naïve Bayes is elegantly Simple and robust, and that is why it is widely used for classifying purposes. It uses the probability theory to classify data. The study also shows that Support Vector Machine is a type of learning system algorithm that is used to make the classification more accurate. To improving the accuracy and speed of Naïve Bayes they used SVM. At the end the researcher confirmed that the correctness and efficiency of the model has been improved for greater expansion in the data set in future.

Simret Solomon [37], Conducted a study on prediction of customer loyalty (Non loyal or Loyal) using the application of data mining in microfinance and built a classification model which supports loan decision making in the organization. In this study a classification model is built based on the loan data obtained from Joshua Multi-Purpose Limited Liability Cooperative (JMPLLC). The study used dataset of 6,447 records and 10 attributes and also used three classification algorithms: Rule Based ZeroR, Decision Tree (J48), and Bayes (Naïve Bayes).

The author performed five experiments, one of it is using Naïve Bayes. In this experiment, classification model building was done using Naïve Bayes classifier with all dataset and attributes. The test model is 10-fold cross validation. It consists of partitioning a dataset into 10 subsets. The experiment resulted with an accuracy of 97.32%. Its ROC area is also above 0.5, which is 0.992. And also, the accuracy level of ZeroR is 55.34% and J48 is 97.31%.

From the results of the experiment, the researcher concluded that the data mining tools and techniques, especially classification techniques, can be effectively applied on the microfinance and financial institutions data in order to generate predictive models with an acceptable level of accuracy.

Sara Worku [46], Conducted a research in the Addis Credit and Saving Institution. The main goal of the research was to develop a classification model that investigates credit analysis in Addis credit and saving institution using data mining techniques. The study employed three classification algorithms to develop the model, namely decision tree, J48, Naïve Bayes and PART rule induction. Cross validation test method was used to validate the model. The researcher followed CRISPDM methodology and used WEKA tool to develop the model. The data was extracted from four branches of the same institution within the same district and the data set contain 4000 records and 7 attributes.

Finally, the author selected PART rule induction that produces optimal accuracy of 99.83% among the nine experiments conducted for the rebalanced dataset with 7 attributes (the study used only seven loan-based attributes).

At the end, the researcher concluded that, the result shows it is possible to extract useful pattern through data mining algorithm that help to make accurate decision for credit risk assessment. But the study conducted through loan characteristics attributes only that is 7 attributes and the dataset is limited in number and governmental affiliated MFI only.

Samuel N. John, Olatunji J. Okesola and Osemwegie Omoruyi [19], conducted a study to investigate the ability of Bayesian classifier towards building credit risk model in banking sector. The study used demographic and material indicators as input variables because of unavailability (privacy reason) of financial data like historical data and borrowers'

characteristics from banks. The study is restricted to the use of demographic and material indicators only. MATLAB were used to partition the sample population to training data and test data for the method, Bayesian algorithm to connect to the partition mode and predict loan request as good or bad.

During the experiment, the study used the total population size of 690 which was partitioned at ratio 34:66 to return a multiclass naïve Bayes model training and test sample data. The model outperformed in classification with 83.3% prediction accuracy.

At the end, the researcher concluded that Naïve Bayesian algorithm has been able to predict credit request as good or bad respectively denoted by 1 or 0 and Naïve Bayesian classifier is good for its speed and memory usage which are really good for simple but not kernel distributions.

Aida Krichene and Abdel moula [38], Conducted a study to tackle the question of probability of default prediction of short-term loans for a Tunisian commercial bank. They used a database of 924 credit records of Tunisian firms granted by a Tunisian commercial bank from 2003 to 2006. The K-Nearest Neighbor classifier algorithm was applied and the results indicated that the best information set is relating to accrual and cash-flow and the good classification rate is in order of 88.63 % (for $k=3$).

The researcher indicated that the outputs of these models also play increasingly important roles in banks' risk management and performance measurement processes. ROC curve is plotted to evaluate the performance of the model. The result shows that the AUC (Area Under Curve) criterion is in order of 95.6%. At the end, the researcher indicated that the study is incomplete in the sense that it didn't show how one can use these results in the implementation of the credit risk or commercial risk in Tunisia banks.

Bolarinwa Akindaini [5], explored the application of some machine learning models in the prediction of mortgage defaults. It basically explored how machine learning methods can be used to classify mortgages into paying, default and prepay. The work examines the machine learning methods: Logistic regression, Naive Bayes and Random forest. The study used the total size of the dataset which is 11250 and about 60% of the randomly selected subset was

used to train the model while the remaining 40% was used as a test set. The overall accuracy of the model was 95%.

At the end, the researcher indicated that the model presented in the work are generic and the results are based on the variables in the dataset. To produce a more robust model, it will be necessary to include variables that gives certain information about the mortgage owner. Such information could include sex, income, occupation and volatility in occupation.

The researcher also indicated that further improvement could be made by clustering the dataset according to this additional variable and build models that are cluster specific. For instance, it will be interesting to build a separate machine learning model for high income earners and low-income earners since chances of default will certainly differ for both groups.

Martina Sandberg [3], Has conducted a research on investigative machine learning model's logistic regression, multilayer perceptron and random forests in the purpose of categorizing between good and bad credit applicants.

The data used for this research were provided by PayEx(it is a financial sector company and offers several payment solutions for channels online and physically) comes from two sources: PxR(internal database of PayEx) and CCP(external database of PayEx). The PxR data contain information about the invoice of the applicant such as the state of it, distribution and expiring date, amount to pay, amount paid, seller information etc. The CCP data include information about the credit decision process for the applicant and contain information such as date of decision, amount requested by the buyer, amount of credit given, reason for rejection. The CCP data set also include information about the applicant such as address, number and age of accounts held, income, payment remarks, gender, number of active claims and verification checks of the information given by the applicant. But the study uses and focus on invoice data and only on reservations which are in the state capture, in total the study used data set for the experiment when the data is merged is 27328 entries with 23 attributes. According to experiment result the author stated that the total dataset is used as train and test sets, 75% and 25% in train and test respectively which resulted in 90% accuracy of the model.

The researcher has also recommended that there are several other algorithms to deal with the problem of unbalanced data that could be tested and also, several other possible variables and algorithms can be tested to find the ones that can do a better job at explaining the difference between defaulting and good customers.

Amir E. Khandani and Adlar J. Kim [2], Studied on applying machine-learning techniques to construct nonlinear nonparametric forecasting models of consumer credit risk. The objective of the study was by combining customer transactions and credit bureau data for a sample of a major commercial bank's customers, they are able to construct out-of-sample forecasts that significantly improve the classification rates of credit-card-holder delinquencies and defaults, with linear regression.

The study presents a method for time-series patterns of estimated delinquency rates from this model over the course of the recent financial crisis suggests that aggregated consumer-credit risk analytics may have important applications in forecasting systemic risk.

During the experiment the study used a data set by combining customer transactions and credit bureau data from January 2005 to April 2009 for a sample of a major commercial bank's customers 7400 entries with 22 attributes, and the model average forecasted/realized delinquencies of 85%.

To the end, the researcher directed that the results are indicative of considerably more powerful models of consumer characteristics that can be developed via different machine-learning techniques, and are exploring further refinements and broader datasets in ongoing research.

Anchal Goyal and Ranpreet Kaur [42], introduced an effective prediction technique that helps the banker to predict the credit risk for customers who have applied for loan. A prototype is described in the study which can be used by the organizations for making the correct or right decision to approve or reject the request for loan of the customers. The study used three different models (SVM Model, Random Forest Network and Tree Model for Genetic Algorithm) and the Ensemble Model, which combines these three models and analyzed the credit risk for optimum results. Real Coded Genetic Algorithms is used to calculate the feature

importance. These features help to predict the credit risk for costumers. K- Fold validation method is used to calculate the robustness of the predictive model.

The data set that the study used for the experiment 615 records which is unfiltered data. The filtered Train Data set file contains 479 records include 13 attributes such as Gender, Marital Status, Number of Dependents, Loan Amount, Credit History and others. All models run on their defaulting parameters and the data is distributed among training and testing set are 70% and 30% respectively for all the models. Performance on the basis of its Accuracy was 80.56% for Random forest ,80.56% for SVM, accuracy of genetic algorithms is 81.25%, decision tree 68.47% and Ensemble model (SVM+ genetic algorithms+ decision tree+ Random forest) that is 77.71%.

The study shows that the above models is useful for banks to decide whether to sanction a loan to the individual or not, for Estimating the probability that an individual would default on their loan.

Lkhagvadorj and Munkhdalai [4], Show that Credit risk prediction is one of the important challenges to the decision-making process for the lending institutions. This researcher applied the deep learning approach for credit scoring based on four datasets. It determines an optimal number of nodes in hidden layers, an effective batch size and epoch number using the grid search method.

The experiment conducted in the study is using four datasets in order to evaluate the performance of proposed algorithm against previous studies. Those datasets retrieved from the UCI repository, namely German 1000 samples with 21 features, Australia 690 samples with 15 features, Japanese 690 samples with 16 features and Taiwan 6000 samples with 23 features, the total of 8380 entries. For German, Australia and Japanese datasets, the deep learning model achieved the accuracy 0.771, 0.8681, and 0.8653, respectively. For Taiwan dataset, the deep learning model showed the best performance. The author stated that Taiwan dataset contains 6000 instances and the sample size is much bigger than other datasets. For this reason, the model outperformed the best by accuracy of 0.92.

The results from the study showed that the proposed algorithm performs very well in a credit scoring application. To the end, the researchers anticipate potential future work in this area that includes using other machine learning model for credit prediction based on the bigger dataset, Especially, bigger sample dataset will give the best performance.

Jozef Zurada and Martin Zurada [6], Conducted a study on banking sectors, and they built a prediction model to predict the customers will pay back the loans or not, through using the techniques of neural network and classification. The researchers focused on the usefulness of the tools such as decision trees and neural networks. In this study the author mainly checks and examining how decision trees and neural networks applied in a credit-risk evaluation.

For the experiment the author used a sample data provided by a money lending institution. The data set contains financial information about 3364 consumers allocated among 13 variables. Out of these 13 variables, there were 12 independent variables (loan and consumer characteristics) and one dependent/target variable (loan default or loan repaid) that they were going to predict. Using this data set, they built a decision tree model, neural network model, logistic regression model, and combined (ensemble) model to predict whether a future applicant will default on a loan. In the first scenario, the author used the original, unbalanced data set containing a total of 3364 customers. This first data set contained 3064 good loans and 300 bad loans. In the second scenario, they created a balanced data set by randomly selecting 300 loans from the 3064 good loans and matching them with the 300 bad loans. This random sampling produced the second data set consisting of 600 cases divided evenly among good and bad loans. In each scenario, they performed three different experiments. In each experiment, they allocated the cases as follows: 60% for training, 20% for validation, and 20% for testing.

In result of the experiment the classification accuracy of good loans was almost perfect (close to 100%), some bad loans were recognized as bad loans, and the three methods and the ensemble method tended to classify a substantial amount of bad loans as good loans. Although the overall classification rate is excellent and averaged about 93%, the average classification accuracy of bad loans, which were underrepresented in the training sample, is relatively low and amounts to an average of about 30% for all the methods across all experiments.

To the end the author concludes that further research should focus on refining the training and testing of the ensemble model, the neural network and other method using various balanced and unbalanced data sets to improve the classification performance.

The above discussed literatures are summarized in the following table.

Table 2.1: - Summary of literatures reviewed on loan risk prediction.

S.NO	Researcher/ Author	Objective	Classification technique	Sample size, accuracy and attribute's	Data set used	Gap (limitation) of the study
1	Paulius and Saulius [40]	To presents an empirical evaluation of SVM-based classifiers applied for credit risk evaluation task	SVM	3266 With 13 attributes, 84%	Commercial bank data set	The study uses only 13 attributes related to specific commercial bank loan product characteristics, and the model accuracy is only 83% it requires improvement.
2	Vimala S.and Sharmili K[39]	To present loan risk Prediction model using bank dataset for commercial banks.	SVM and Navies Bayes	850 With 21 attributes, 81% Accuracy	UCI data set	The model performance in terms of accuracy is not good it requires to improve and the data set used is limited in number and the data set used is very specific to commercial banks not include MFIs loan and borrowers characteristic.

3	Aida Krichene and Abdelmoula [38]	To design model for prediction of probability of default in short term loans for a Tunisian commercial bank.	KNN	924 With 11 attributes, 88.63% Accuracy	Tunisia n bank	The study used only 11 loan characteristic attributes specific to Tunisian commercial bank one loan product with only take 924 records and also limited to test other ML algorithms by scope only one algorithm.
4	Simret Solomon [37]	To design a classification model for prediction of customer loyalty in microfinance. (loyal or not loyal)	Bayes (Naïve Bayes). Decision Tree (J48)	6447 With 10 attributes, 97.32% and 97.31%	JMPLL C data set	The study is conducted only one microfinance not inclusive to other MFIs and the attributes and the data set of the study is very limited in number and categories and also focused in one product characteristics.
5	Sara Worku [46]	To design a model to investigate credit analysis in Addis credit and saving.	Bayes (Naïve Bayes). Decision	4000 With 7 attributes, 99.83% accuracy	Addis credit and saving	The study is done for microfinance but the attributes used for the study is very limited only 7 and about only one loan product characteristics, the data set also very limited in number and not inclusive other microfinance institutes.

6	Samuel, Olatunji . Okesola and Osemwegie [19]	To investigate the ability of Naïve Bayes towards building credit risk model in banking sector.	Naïve Bayes	690 sample 83.3 accuracy	Bank dataset	The model performance of the model is needs to improve and the data set used also very limited in number and the model build for specific loan products only.
7	Bolarinwa Akindaini [5]	To examine and explore the application of some machine learning models in the prediction of mortgage defaults.	Logistic regression, Naive Bayes and Random forest.	11250 sample and 95% accuracy.	Credit card data sets	The study indicates it limitations by itself in terms of including loan borrowers characteristics to build more inclusive model.in addition the data set used in the study also limited in scope and in number.
8	Martina Sandberg [3]	To investigate machine learning algorithms in the purpose of categorize between good and bad credit applicants.	Logistic regression, multilayer perceptron and random forests	27328 entries with 23 attributes and 90% accuracy	PayEx data sets	The study concern on predict online credit payment solutions by taking only the card owner last transaction history, it shows good accuracy but not take the loan and card owners specific characteristics.

9	Amir E. Khandani and Adlar J [2]	To present nonlinear nonparametric forecasting models of consumer credit risk for commercial banks by applying machine-learning techniques.	Logistic regression	7400 entries with 22 attributes and 85% accuracy	Commercial banks data sets	The study used 7400 entries and from only one commercial bank loan product also the result of the model still needs to improve.
10	Anchal Goyal and Ranpreet Kaur [42]	To introduces an effective prediction technique to predict the credit risk for banks.	Ensemble Learning	479 records include 13 attributes and 77.71% accuracy	Commercial banks data sets	The data set used in this study is only 479 records with only 13 attributes and the result of the model accuracy is 77% so it needs to improve.
11	Lkhagvad orj and Munkhdalai [4]	To show and explore that credit risk prediction model importance for lending institution by designing a model.	deep learning	Use four data sets 6000,1000,690,690, and 0.771, 0.8681,0.8653 ,0.92, accuracy respectively.	UCI data sets	The data set the study used for each model is not big and the accuracy of the model by itself needs to improve.

1	Jozef	To build a model	decision	3364 with 13	Comme	The data set they use is
2	Zurada and Martin Zurada [6]	that analysis a customer loyalty in loan.	trees and neural networks	attributes 93% of accuracy	rcial banks datasets	limited only 900 records and also the model build for banks loan with limited number of attributes and inclusive to MFIs.

2.6. Gap Analysis

It is known that a number of financial institutions are in action today throughout the world. Financial institution include banks and microfinance institutions has their own product design and service delivery strategies [25]. Financial institution specially microfinance institutions practice a different challenge and different loan borrowers' characteristics related to identifying good borrowers for loan approvals problem as stated in [50] [58]. So, one can say that the problem of analysis of loan risk and identifying good and bad borrowers varies from one financial institution over another because of their different types of product design and strategies of loan disbursement.

Although a few studies have explored in the area of loan risk assessment and prediction specially most of them are in bank sectors. A direct implementation of those studies output is not practical for Ethiopian microfinance institutions. The loan borrower's characteristics and the loan product design differ from country to country [50] [58]. This research conducted on Ethiopian MFIs (selected seven Ethiopian MFIs: - Oromia, Aggar, Wosassa, Pease, Vision fund, Nisir and Harbu MFIs).

The study Attempt to review different literatures related to loan risk prediction for different financial institutions like banks and microfinance. In addition, existing studies used different approach such as Different attributes in terms of loan, borrowers and business characteristics, different data sets and preprocessing activities, and different method for their studies, because

of most financial institutions follow different way of lending loan, strategies and MFIs structure throughout the world [30].

Studies like [5][2][3], stated in related works above, suggest that the results will improve and more powerful model could be developed by including different attributes related to borrower's characteristics and different data sets. In addition to this researchers [2] [4]- [6] [39] [40], indicated that further improvement can be done and tested through, clustering the dataset according to additional attributes and build models that are cluster specific by categorizing high income earners and low-income earners (loan collateral or business income attributes) since chances of default will certainly differ for both groups and by relatively broader datasets.

In this study the problem of loan risk prediction was addressed using borrower's characteristics which are significant in real world loan risk assessment practice but not included in previous studies like [46] [37] to make better prediction and more relevant and practical in all MFIs. Studies like [50], conducted a study on Ethiopian microfinance loan borrowers and identified loan characteristics, business characteristics and loan borrowers' characteristics as attributes for loan risk assessment and identification of good and bad borrowers. Such attributes include monthly income earns, occupation, business type, location of collateral, business location of borrower's, MFIs types and academic status of borrowers are included in this study. In addition to that, the study used different data sets (not used before for related works) with balanced and relatively large sample data sets with adequate representation of all geographical locations in Ethiopia by taking the data from MFI's branches in different cities unlike to the prior studies. In general, as stated above this research uses additional new attribute that include borrowers specific, borrower's current business experience in year, microfinance type, location of collateral or business in categories, yearly business income, total years of experience in any business and loan product type, and also the study used a new data sets (from selected seven Ethiopian microfinance) that is not used before for related study for loan risk prediction in Ethiopian MFI. The model to be developed will be relevant and practical for Microfinance Institutions in Ethiopia.

CHAPTER THREE

RESEARCH METHOD AND TECHNIQUES

3.1 Introduction

This chapter discusses the design and methodology of the experiment that will be carried out to answer the research question of this study. Research methodology is the general principle that guides the research. In order to conduct a good research, a well-defined approach and principle has to be followed. This study follows the experimental type of research. It is a collection of research design which use manipulation and controlled testing to understand causal (cause/effect) relationships and to study the relationship between one variable and another. The design of this experiment follows the CRISP-DM process outlined in figure 3.1.

For proper understanding of the problem under investigation and successful completion of this study, relevant literatures such as books, journals, magazines, conference papers, manuals, and resources from internet, particularly MFI loan manuals are reviewed for achieving the study objectives.

3.2 The Cross-Industry Standard Process for Data Mining (CRISP-DM)

The aim of this study is to build a predictive model that could help the organization in predicting active and defaulter loan borrowers in early stage before loan dispersing and to secure their portfolio at risk precisely. New customers' status that supports loan decisions. To this end, this study followed the CRISPDM (Cross-Industry-Standard-Process for Data Mining) process cycle. The cross industry standard process for data mining (CRISP-DM) usually shortened as CRISP DM is a framework for guiding and recording data mining tasks. It is a model that consists of phases that are followed to solve data mining problems [70].

The reason why CRISP-DM methodology use because CRISP has feedback mechanism SEMMA have no feedback mechanism, SEMMA was developed with a specific data mining software package (Enterprise Miner), rather than designed to be applicable with a broader range of data mining tools and the general business environment like CRISP-DM, in addition to this. It is easy to use, it is the most widely used model that experts and non-experts use in solving data mining question. And also it is a popular methodology adopted by many Data

Miners because when followed critically there is a higher chance in succeeding in the DM project [69].

The CRISP-DM process model starts with the project's goal definition that is included in the first phase that is Business Understanding, then transformed into a specific data mining problem. During the Data Understanding phase hypotheses for hidden information regarding the data mining project goal are formed based on experience and qualified assumptions. In the Data Preparation phase the researcher collects the relevant data and prepares it for the actual data mining task. This includes the data preprocessing such as data reduction and filtering. In the Modeling phase a data mining workflow is constructed to find the desired parameter settings for the selected algorithms and to execute the data mining task on the preprocessed data [69].

Within the subsequent Evaluation phase the trained model is tested against real data sets within a production scenario and the data mining results are assessed according to the underlying business objectives [69].

3.2.1 Business Understanding

This phase is also known as Problem understanding and entails the processes used to comprehend the opportunities and business purposes of the company. Ethiopian microfinance sector is characterized by its rapid growth, an aggressive drive to achieve scale, a broad geographic coverage, a dominance of government backed MFIs, an emphasis on rural households, the promotion of both credit and savings products, a strong focus on sustainability and by the fact that the sector is Ethiopian owned and driven [10].

From the interviews and discussions made with senior managers of the MFI institutions the existing loan risk analysis and loan granted is dependent on loan officer views of borrowers and loan applicant commitment form of borrowers, it requires significant improvement to minimize the loan risk and identify defaulter borrower's in early stage before loan disbursement.

Almost all selected microfinance institutions are at risk regarding high rate of default/delinquency by their clients; which are most of microfinance institutions are not

achieving the internationally accepted standard portfolio at risk of 5% [25], which is a cause for concern because of its consequences on MFIs businesses, individuals, and the economy of Ethiopia at large.

Therefore, his research aims to develop loan risk prediction model using machine learning algorithms for MFIs in Ethiopia. The research focus on collecting and analysis of loan application data to accurate risk assessment and to manage portfolio at risk ratios precisely

3.2.2 Data Understanding

This is the next phase of the CRISP-DM methodology. For this study, the data is collected from selected seven microfinance institution in Ethiopia. As discussed with domain experts and senior managers of MFIs and also as stated in [50], (it is a social science thesis study that concerns on Ethiopian microfinance institutes and also the study specifies the main reasons (attributes) of borrower's loan delinquency), the data from MFIs contains borrowers' characteristics, loan characteristics and business characteristics attributes. The data contained 37380 loan borrowers' records with 18 attributes one of the attributes is dependent field representing either a default or not a default field; the data collected from all selected MFIs as of December 31 2018. No single prospective customer is contacted more than once. and Single (Marital status) respectively.

3.2.3 Data Preparation

Taking into consideration the organizational policies This study is conducted not to include any private or confidential data of any loan borrower (customer of MFIs). Names, phone no and specific addresses is excluded from the data set. In order to consider the research to be ethical, anonymity and confidentiality administrative permission is obtained. Permission is obtained from the chef executive manager, branch manager, loan officer and the information technology head before any information is released.

This phase is also known as data filtering and involves the process of organizing the data for mining. The following steps are completed in order to filter the data to be used in the DM process. These are data cleaning, data transformation data quality assurance, combining data sets, choosing part of the data as subgroup, combining rows, developing new columns,

arranging the data to be used in the modeling, taking care of problematic figures (blank or missing) and dividing into training and test data sets

3.2.4 Predictive Modelling

Predictive modeling is the general concept of building a model that is capable of making predictions. Typically, such a model includes a machine learning algorithm that learns certain properties from a training dataset in order to make those predictions [53].

Predictive modelling is a name given to a collection of mathematical techniques or models that helps in finding a mathematical relationship between a target or dependent variables and the predictor or independent variables [54]. It helps in predicting the probability of an outcome when a set of independent variables passes through the model. SVM, KNN, Naïve Bayes and logistic regression models can be used for prediction purposes.

3.2.4.1 Reason for The Chosen Algorithms

The principal goal of this research is to analyze the existing MFIs loan risk in Ethiopia and predict loan risk using computational algorithms. With this in mind, the study aimed at identifying machine learning technique which is better in predicting loan risk. Throughout this work different machine learning algorithms were explored and effective ones were used to find patterns in the data. The performance of each machine learning models was explored and analyzed in previous related works. Then based on successfulness in making predictions and stability in their performance best performing algorithms were selected. The machine learning techniques selected for this thesis were SVM, KNN, Navies Bayes and Logistic regression. Each one is selected based on their advantages and past performance seen in other research.

In different literatures, it has been reported that the widely used classifier algorithms for prediction and classification are KNN like in [16], [17], [23], SVM like in [18], [21], [23], [24], Logistic Regression like in [15] and Naive Bayes like in [19], [20], [24]. The study was use the above four different algorithms to build four different models for this loan risk prediction and classification model.

Naïve Bayes is selected due to the following reasons; it is easy to implement, Naïve Bayes classifiers can be trained quickly [19], classification process is quick compared to other models

[20] [24], it can handle a large and discrete amount of data, it is not sensitive to irrelevant features [55].

Logistic regression which is also called logistic model or logit regression is a predictive analysis. It takes independent features and returns output as categorical output. The probability of occurrence of a categorical output can also be found by logistic regression model by fitting the features in the logistic curve [35]. Logistic Regression is included in these work because; it is easy to implement and no linear relationship between independent and dependent variable [56], multiple explanatory variables can be used, no confounding effects because logistic regression allows quantified values for strength of association between explanatory variables and less prone to over-fitting due to simplicity and low variance [15].

K-nearest Neighbor is selected because; it is difficult to imagine a simpler technique than KNN, where data is classified simply based on its nearest neighbor (or neighbors) in a given training set [23]. Learning does not require making any assumption about the characteristics of the concepts [16] and Complex concepts can be learned by local approximation using simple procedures [23].

SVM is selected to this study because; it is established on the structural risk minimization principle, which seeks to minimize an upper bound of generalization error, and is shown to be very resistant to the over-fitting problem [18] [24]; it uses the kernel trick, so can build in expert knowledge about the problem via engineering the kernel [24]. SVM model is a linearly constrained quadratic program so that the solution of SVM is always globally optimal, while other models may tend to fall into a local optimal solution [21].

3.2.5. Model Evaluation

This section presents the evaluation methods that was employed to evaluate the performance of the loan risk prediction models. One of the popular performance evaluation methods in machine learning, data mining, artificial intelligence and statistics is confusion matrix. In order to quantify the performance of a problem that contains two classes, confusion matrix is usually used [22]. The study used Confusion matrix as evaluation methods because it comprises of

evidence about actual and predicted classification performed by the proposed prediction and classification model.

3.2.5.1 Confusion Matrix

The confusion matrix is used to measure the performance of two class problem for the given data set. The right diagonal elements TP (true positive) and TN (true negative) correctly classify Instances as well as FP (false positive) and FN (false negative) incorrectly classify Instances [22].

TN = the number of incorrect classifications that an instance is Negative.

FP = the number of incorrect classifications that an instance is positive.

FN = the number of correct classifications that an instance is Negative.

TP = the number of correct classifications that an instance is Positive.

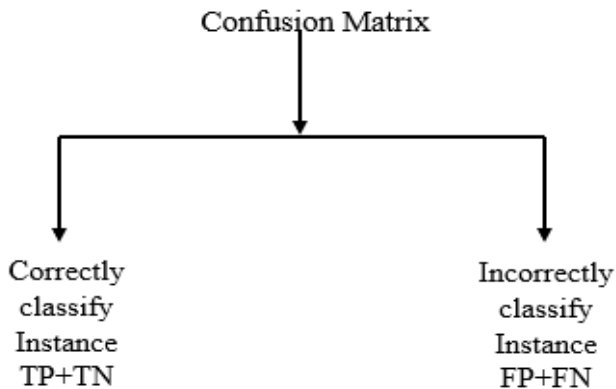


Figure 3:2 Example of Confusion Matrix

	Predicted	
Actual	Yes	No
Yes	TP	FN
No	FP	TN

Table 3. 2: Example of Confusion Matrix

Total number of instances = Correctly classified instance + Incorrectly classified instance

Correctly classified instance = TP + TN.

Incorrectly classified instance = FP + FN

Calculate Value TPR, TNR, FPR, and FNR One can calculate the value of true positive rate, true negative rate, false positive rate and false negative rate by methods shown below [22].

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \dots\dots\dots(3.1)$$

$$\text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}} \dots\dots\dots (3.2)$$

$$\text{FPR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \dots\dots\dots (3.3)$$

$$\text{FNR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \dots\dots\dots (3.4)$$

Precision: -Precision is the ratio of modules correctly classified to the number of entire modules classified fault-prone. It is proportion of units correctly predicted as faulty [22].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \dots\dots\dots (3.5)$$

Accuracy: -Accuracy is defined as the ratio of correctly classified instances to total number of instances [22].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \dots\dots\dots (3.6)$$

3.3 Privacy and Confidentiality of Borrowers Data

This study is conducted not to include any private or confidential data of any loan borrower (customer of MFIs). Names, phone no and specific addresses is excluded from the data set. In order to consider the research to be ethical, anonymity and confidentiality administrative

permission is obtained. Permission is obtained from the chief executive manager, branch manager, loan officer and the information technology head before any information is released.

3.4 Summary

This chapter discussed the methodology used for conducting this study. The chosen research method, model evaluation techniques and, the CRISP-DM methodology was also explained in general terms and in the light of how it was used and guide this study. The loan borrower's data ethical issues are also discussed in terms of privacy and confidentiality.

CHAPTER FOUR

DATA PREPARATION

Data preparation is some of the primary tasks that highly determine the data mining results. The model built mainly depends on how thoroughly and carefully the necessary data is obtained, analyzed and preprocessed. Hence the next subsequent sections present the data and the essential preprocessing tasks performed.

4.1 Data Preprocessing

The data preprocessing refers preparing the dataset in the form that it is ready to Machine learning task. In this study the processes applied include data cleaning, parsing, data selection and aggregating on the extracted data in order to make the data more suitable for the experiment to improve the overall machine learning task.

4.1.1 Data collection

For this study, all the data was collected from selected seven microfinance institution in Ethiopia out of thirty-five MFIs based on their outstanding balance, electronic data availability, the scale of operation (gross loan portfolio), and on their ranks, in order to avoid bias in the research finding. Three of them are from NGO affiliated MFIs (Vision Fund, Peace and Wosasa MFIs), and one from government affiliated MFIs (Oromia MFI), the rest three are from fully commercial MFIs (Aggar, Niser and Harbu MFIs). The data included new attributes which are not considered in prior studies that include borrower's specific data such as borrower's current business experience in year, microfinance type, location of collateral or

business in categories, yearly business income, total years of experience in any business and loan product type. The dataset also included additional attributes such as loan amount granted, purpose of loan, collateral, age, history of payment and others. Records contain a dependent field representing either a “Defaulter” or “Active” field.

4.1.2 Data Cleaning

Among the total dataset extracted, 347 of them have missed values for attributes such as education status and business yearly income. Instead of filling values on these attributes, it was found easier and more logical to remove the records that make up 0.92% of the dataset. As a result, the remaining 99.08% of the original dataset which amounted to 37,380 records were kept for further processing.

4.1.3 Method of Data Quality Assurance

The datasets were checked for completeness and correctness of the required attributes and integrity before analysis and prediction. The study explored different non-machine learning studies related to loan risk prediction using financial data and risk assessment strategies in Ethiopia’s microfinance institutions. This helped us to ensure whether the required attributes (column names) are complete and adequate for prediction of loan risk for this study. The data from MFIs contained required loan and borrowers’ details.

4.1.4 Imbalance Data and Splitting the Data Set

An imbalance dataset is such a case where there is major difference in the number of classification categories [57]. In our data set domain, the classification categories consist of “Defaulter” and “Active”, where the number of “Active” cases outnumber the number of “Defaulter” cases. Almost 62 percent of the datasets are non-defaulters (Active). In such a situation a model becomes more inclined to the majority class and cannot properly identify the minority class. To solve this issue, there is two possibilities either we can over sample the minority class or under sample the majority class. But under sampling the majority class will act as a difficulty in properly understanding the trends in our independent attributes. Furthermore, only over sampling the minority class will also not solve this as the techniques lying behind the over sampling will also matter greatly. Thus, in such a scenario the study used Synthetic Minority Oversampling Technique (SMOTE) and up sampling. This technique uses

for both oversampling and under sampling [57]. Synthetic instances of the minority class are created to reduce the margin between the majority and minority class [57]. For the model the study used up sampling to increase the minority class and keep equal number of defaulters and non-defaulters. But we have to mention that up sampling was only applied on the training set keeping the test set pure and untouched. And therefore, this helped us to properly classify the borrowers keeping the model aware of both the output classes.

For the purpose of checking the performance of any machine learning model in an effective manner, splitting the dataset is a fundamental task. It helps to prevent over fitting by evaluating the performance of the model on a portion of the dataset upon which the model has not been trained. In most empirical studies like [62] [63] [64] it is shown that the data set is split in to 80:20 ratios which has become most common in many other studies. Therefore, this study used 80:20 train-test split ratio for the supervised model. This has been done using R studio “split_test_train” function from ISRL library. This means that 80 percent of the entire dataset was used to train the model and the remaining 20 percent was used to evaluate the performance of the model.

4.2 Data Transformation

Transformations and aggregations of the data is necessary because it minimize the variations of the attribute values in some of the fields and also to make results more expressive and simply interpretable.

In the data set, age attribute is varying in values and it is transformed in to more aggregated values “young, “adult” and old” for the different age groups determined based on experiences and consultation with the domain experts. The dataset doesn’t contain any instance with age below 18 because the minimum age for granting a loan in MFIs is 18 and above. So, there was no need to include another category under age 18. Hence young is considered to be between 18-25, while adult is considered to be between 26-55 and old age was considered to be those with ages 55 and above.

Similarly, to make it easier for discussions and interpretations of the result, other attributes like loan amount, city and education status were generalized in to categorical values. Loan amounts

below 100,000 are categorized as Low, those which are between 100,000 and 500,000 were categorized as medium and those loan sizes above 500,000 were categorized as high.

City (city of the business area or the collateral location of the borrowers) were categorized as “region” if the city is capital city for the regions and “zone city “if the city is capital for the zone and the rest city categorized as “woreda”. Education status of borrowers was also categorized as graduated, high school complete, elementary level and not educated. Outstanding principal balance attributes refers to unpaid Remaining balance from the total loan amount. It was transformed in to more aggregated values, that is” full payment unpaid”, “half and above unpaid”, “less than half unpaid”.

4.3 Attribute Selection

Features Selection is one of the core concepts in machine learning which hugely impacts the performance of your model [52]. The data variables that you use to train your machine learning models have a huge influence on the performance you can achieve [52]. Variables selection and Data cleaning should be the first and most important step of your model designing. Variables Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in [52].

Feature selection is a technique that removes noisy and redundant features to improve the accuracy and generalizability of a prediction model. Although feature selection is important, it adds yet another step to the process of building a bug prediction model and increases its complexity [51].

As stated and suggested in [50] and through made discussion with senior domain experts in MFIs we identify, there are a few borrowers’ characteristic specific to Ethiopia’s MFIs customers and loan characteristics factor that may be considered as the risk factors for loan risk prediction in MFI.

This section provides the description of the risk factors used to predict the loan risk in Ethiopia MFIs. The Borrowers characteristics like Age, Education status, Gender and having other sources of income, are the variables that can influence loan repayment performance of the borrowers [50].

The loan related characteristics include the loan size, loan utilization, credit timeless, repayment suit, collateral, peer monitoring and the credit timeless factor also significant in influencing repayment performance [50] [25].

Business characteristic related variables are analyzed, thus, having high and low market demand for the end products, Technology adoption, business experience, keeping a book of records, supervision and training are also found to be factors influencing loan repayment performance of borrowers [50].

Attribute selection from the data set was done based on the objective of the study at hand. Hence the account number, customer’s names and branch code, loan officer code attributes are removed in order to reduce the data to only most important ones; this would minimize the effort required for further processing.

4.4 Description of the Data Set

For this study, the data is collected from selected seven microfinance institution in Ethiopia. As stated in [50] (it is a social science thesis study that concerns on Ethiopian microfinance institutes and also the study specifies the main reasons (attributes) of borrower’s loan delinquency), the data from MFIs contains borrowers’ characteristics, loan characteristics and business characteristics attributes. The data contained 37380 loan borrowers’ records with 18 attributes one of the attributes is dependent field representing either a default or not a default field; the data collected from all selected MFIs as of December 31 2018. No single prospective customer is contacted more than once. The attributes, description and categorical values explored for the study are described in Table 4.1.

Table 4. 1: Attributes of data set used for MFIs loan risk prediction.

No.	Attributes	Description
1	City status	Capital city, zone or wereda
2	Education status	Education status of borrower
3	sex	Gender of the loan borrower’s
4	Total year of experience in any business	Total years of any business experience

5	Year since current business established.	Number of years in current business
6	Business Sector	Business type like trade, manufacturing ...
7	Interest rate	Interest rate either flat or declining.
8	Percentage of collateral ownership	Borrowers percentage of ownership from guarantee
9	Business yearly Earning	Yearly income
10	Approved Loan amount	Total taken loan
11	Loan cycle	Repetition of loan taken like 1 st or 2 nd time
12	Loan term in month	Loan duration in month
13	Age	Age of the loan borrower's
14	Outstanding principal balance	Current balance (Remaining unpaid balance (payment history))
15	Product type	Loan product type
16	MFIs type	MFIs (NGO affiliated, fully commercial and government affiliated)
17	Loan status	Active or defaulter
18	hours worked per week	Loan borrower's working hours per week

4.5 Data Transformation

From the source of the dataset, the names and values of the attributes have been changed to some generic symbols for the sake of simplicity for experiment and to have a more accurate representation of the variables. The values in the attributes like loan status attributes changed to AA , Sex attributes changed to AB and Age attributes changed to AC, Education attributes changed to AD, Total Years Of Business Experience attributes changed to AE, Year Since Current Business Established attributes changed to AF, business Sector attributes changed to AG, Percentage of collateral Ownership attributes changes to AH, Business Yearly Earnings attributes changes to AI, Hours Worked Per Week attributes changes to AJ, loan product attributes changes to AK, loan cycle attributes changes to AL, approved loan amount attributes

changes to AM, loan term in month attributes changes to AN, outstanding principal balance attributes changes to AO, City attributes changes to AP, MFIs Type attributes changes to AQ and interest type attributes changes to AR.

The dataset in this experiment and analysis contains categorical values that are transformed to binary values or factors 1s and 0s. Sex variable having values 'M' changed to '1' that representing male and 'F' changed to value '2'.

Similarly, education attributes changed to '1' for graduate, '2' for high school, '3' for primary and '4' for not educated attributes, and also for service attributes value changed to '1' for manufacturing, '2' for service, '3' for trade, and '4' for other services.

Product type attribute changed the variables value to '1' for micro loan, '2' for small loan, '3' for WEDEP (Women development program) loan, '4' for general loan, '5' for agricultural loan and '6' for others loan products. And city status attributes variables changed to '1' for zone, '2' for region and '3' for wereda. Also, outstanding principal balance attributes value changed to '1' for more than half of the loan is remaining to repay, '2' for less than half remain, and '3' for total amount is not repay.

Interest rate attribute has two values declining and flat, the value of declining variable changed to '1' and for flat changed to '2'. MFIs type attributes in the original data set representing through NGO affiliated, government affiliated and fully commercial MFIs value, these values changed to '1', '2', '3' values respectively. Below is the short summary of the changed variable names and values which are used throughout the experiment and analysis.

Table 4.2: Variables and transformed data types

Attributes	Transformed attributes values	Original data type	Transformed data type
Loan status	AA	Character	Binary
Sex	AB	Character	Binary
Age	AC	Number	Factor
Education	AD	character	Factor

Total Years Of any Business Experience	AE	Numeric	Not changed
Year Since Current Business Established	AF	Numeric	Not changed
Business Sector	AG	character	factor
Percentage of collateral Ownership	AH	Numeric	Not changed
Business Yearly Earnings	AI	Numeric	Not changed
Hours Worked Per Week	AJ	Numeric	Not changed
Product type	AK	Character	Factor
loan cycle	AL	Numeric	Not changed
approved loan amount	AM	Numeric	Not changed
loan term in month	AN	Numeric	Not changed
Outstanding principal balance	AO	Numeric	Factor
City status	AP	character	Factor
MFIs type	AQ	Character	Factor
Interest rate	AR	character	Binary

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	Active	2	32	1	8	1	1	50	1200000	56	3	2	100000	12	1	2	2	
2	Active	2	50	2	9	9	1	50	500000	84	3	4	150000	12	1	2	2	
3	Active	1	25	1	2	2	1	50	450000	48	3	2	700000	12	1	2	2	
4	Active	2	38	1	12	6	3	50	190000	54	3	2	150000	12	1	2	2	
5	Active	2	30	2	6	4	3	50	60000	78	3	2	5000	12	2	2	2	
6	Active	1	40	3	11	8	3	50	500000	72	3	1	10000	12	2	2	2	
7	Active	1	43	2	20	10	3	100	300000	54	3	2	25000	12	1	2	2	
8	Active	1	36	2	8	8	1	100	25000	12	3	2	13000	12	1	2	2	
9	Active	2	47	2	12	12	4	100	2000000	12	3	1	100000	12	1	2	2	
10	Active	1	42	2	6	10	1	100	300000	48	3	3	20000	12	1	2	2	
11	Active	2	40	2	2	3	1	100	150000	48	3	2	8000	12	2	2	2	

Fig 4.1 Transformed dataset

CHAPTER FIVE

EXPERIMENT AND DISCUSSION OF RESULTS

5.1. Introduction

This chapter discusses how the experiment was carried out, based on the steps mentioned in chapter three. It presents how the machine learning models were created using SVM, Logistic regression, Navies Bayes and KNN, the major experiments run, interpretations and their performance evaluations of the prediction model. Several subsequent tasks in the experiment are done using R studio tools.

In chapter four, all the preprocessing activities performed on the dataset and some of the major tasks performed were presented. This section focuses on presenting summary of the major experiments made in the process of arriving at the optimal model to achieve the objective set in chapter One.

Once the necessary data is passed through the preprocessing activities as described in the earlier sections, it is then loaded to the R studio for the model building required. The preprocessed data is converted to csv format that is suitable for R studio.

Series of Experiments are conducted based on which algorithms prediction models with varying accuracies, sizes and precisions are obtained. This section also presents several

activities done related to running and evaluating model building experiments, selecting the best and appropriate model, and providing explanations on the selected model.

5.2 Data Visualization

Data Visualization is a vital tool that can unearth possible crucial insights from data. If the results of an analysis are not visualized properly, it will not be communicated effectively to the desired audience [59].

A visual analysis is conducted on the dataset to have an idea of the possible relationships between the variables and observe any visible effect of each attributes. R programming offers a set of inbuilt functions and libraries to build visualizations and present data.

This section details on distribution of continuous Variables, this six continuous variables Age, approved loan amount, income, business experience year, current business experience year and loan cycle is observed initially to have a sense of the nature of the dataset. The below fig 5.1 shows the Age frequency of loan borrowers in MFIs, most of the borrowers are in Age range between 20 to 40 and followed by the Age range of 40 to 60. Also, below fig 5.2 shows that most of disbursed loan amount to borrowers in MFIs is less than five million birrs. The below fig 5.3 shows most of the MFIs borrower’s collateral ownership is 100 present and followed by 50 present and a few with 75 present ownerships. And the below fig 5.5 shows that most of loan term in MFIs are 12 months followed by 24 months and a few with 18 months.

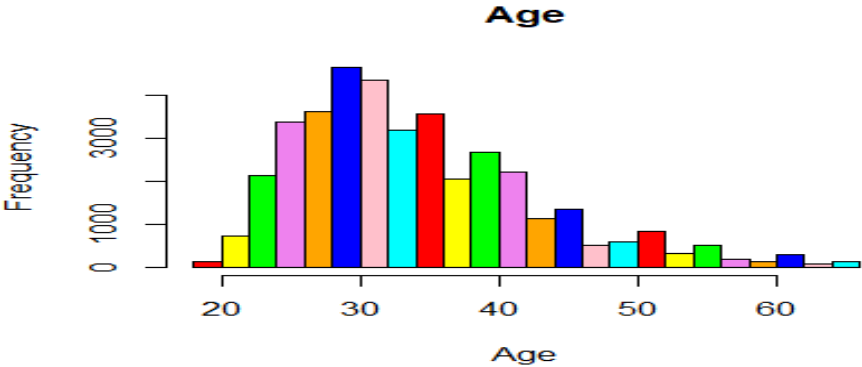


Fig 5.1 Frequency distribution of Age

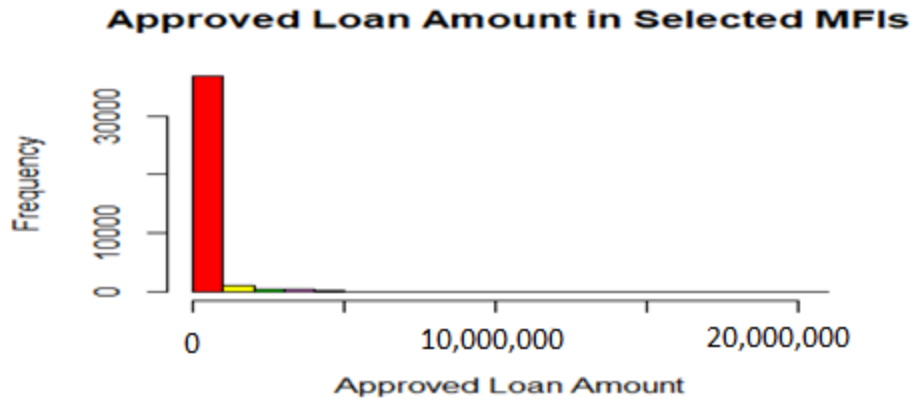


Fig 5.2 Approved Loan Amount

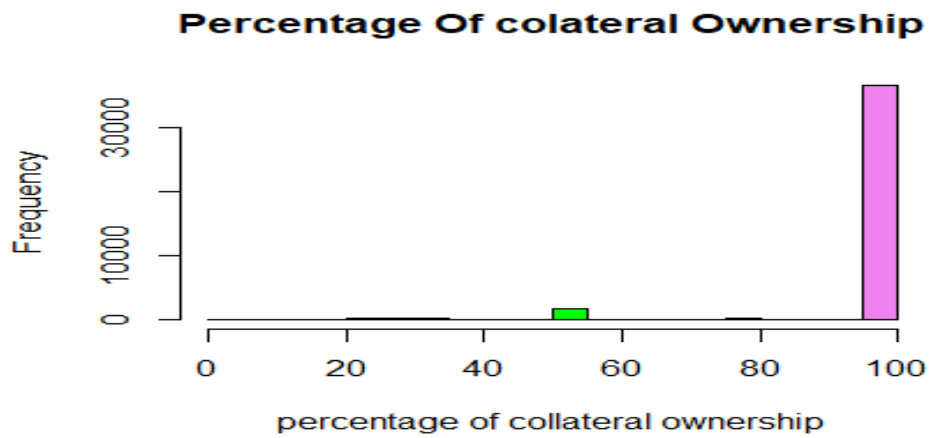


Fig 5.3 Percentage of Collateral Ownership

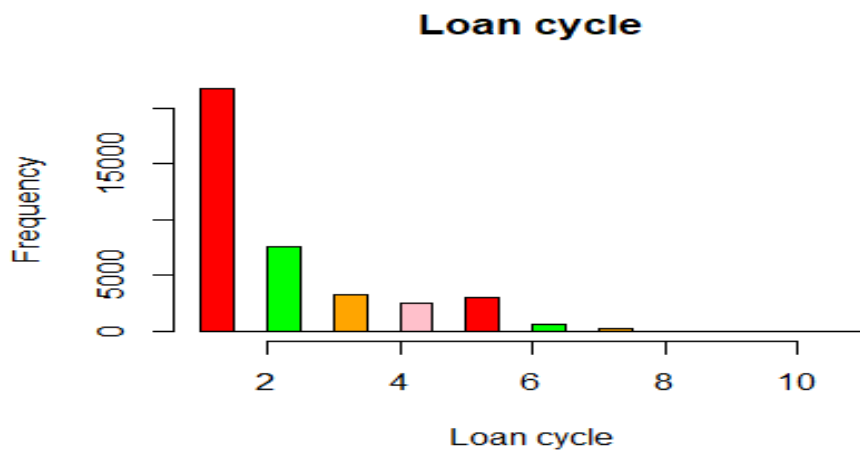


Fig 5.4 Loan Cycle

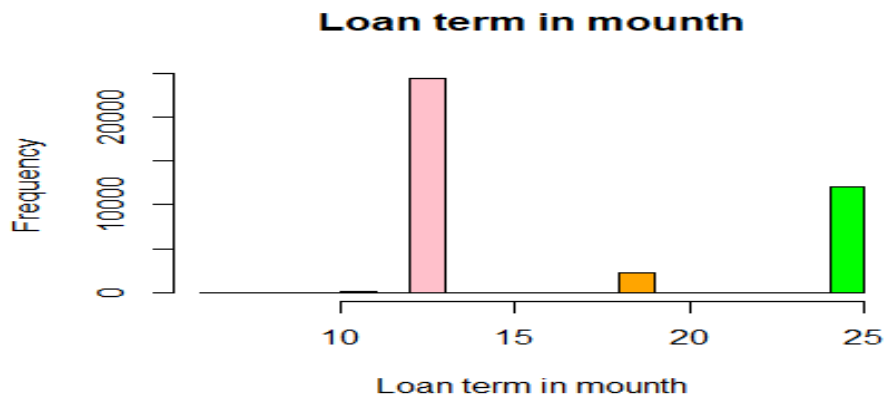


Fig 5.5 Loan term in month

Individual attributes effect on loan status, Plots of some individual attributes is done to inspect visually whether the attributes influenced on loan status. Below are plots of the attributes that appear to effect whether a loan is active or defaulter with loan cycle, loan term and loan product.

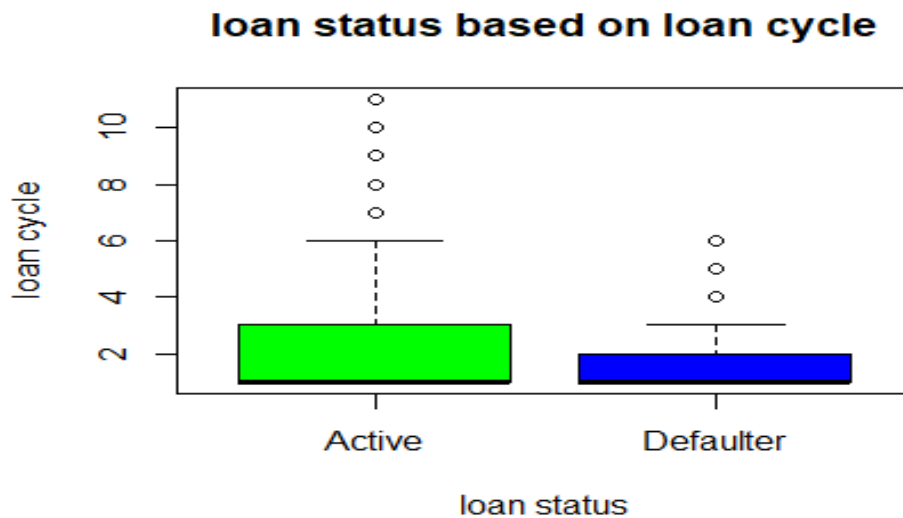


Fig. 5.6: Plot of loan status vs loan cycle

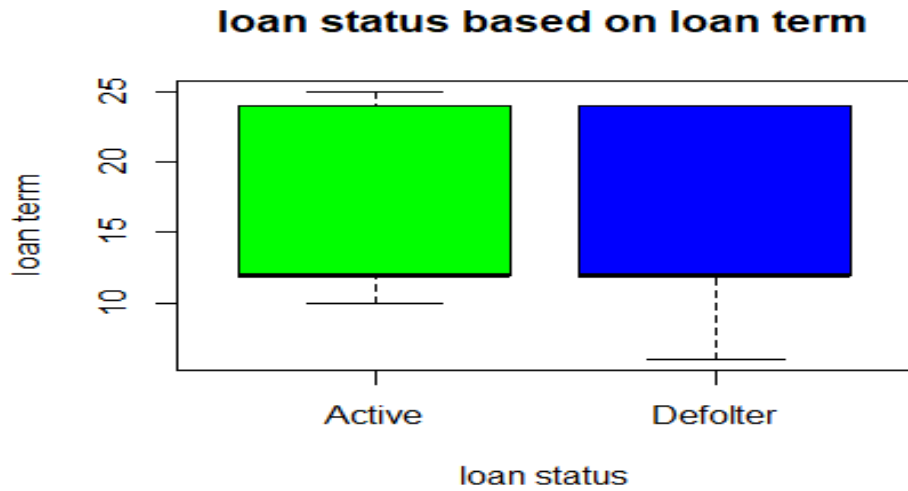


Fig. 5.7: Plot of loan status vs loan term

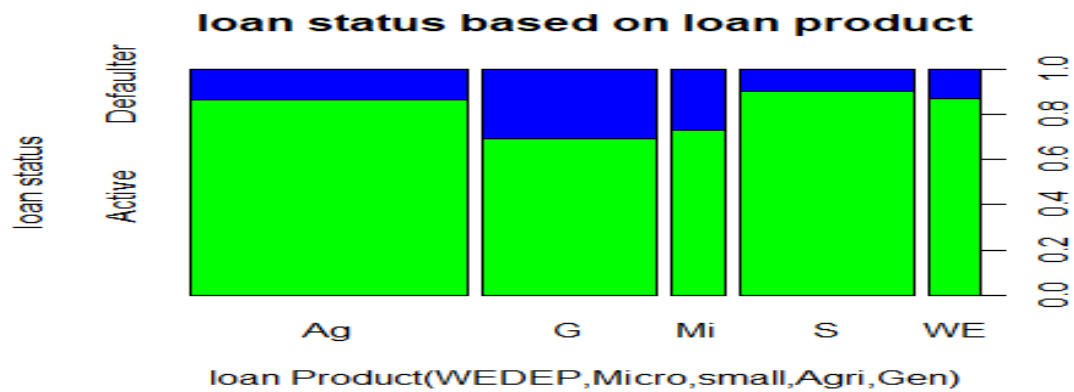


Fig. 5.8: Plot of loan status vs loan product

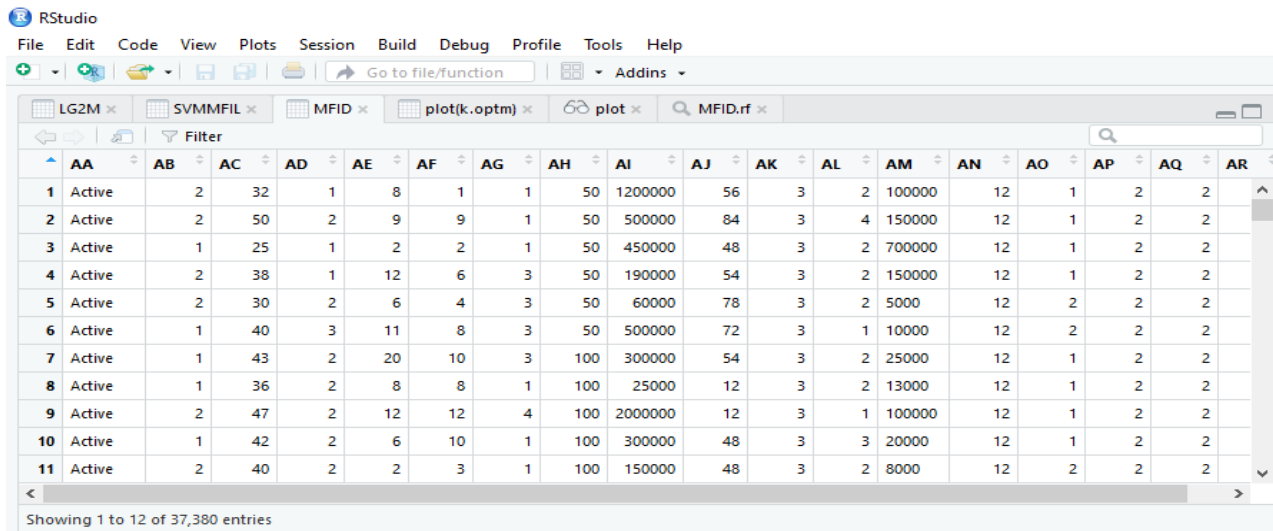
5.3 KNN Modeling

This subsection deals with how the KNN prediction model is developed and how the information gain was calculated. In this experiment a Prediction model building is done using KNN algorithm.

5.3.1 Experiment One

The data set consists of 37380 observations and 18 attributes with one dependent attribute (Loan status) and all of the predictor (independent) attributes were used in this experiment.

Here's how the data set looks like:



The screenshot shows the RStudio interface with a data frame snapshot. The data frame has 18 columns labeled AA through AR and 11 rows of data. The first column (AA) contains the word 'Active' for all rows. The other columns contain numerical values. The status bar at the bottom indicates 'Showing 1 to 12 of 37,380 entries'.

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	Active	2	32	1	8	1	1	50	1200000	56	3	2	100000	12	1	2	2	
2	Active	2	50	2	9	9	1	50	500000	84	3	4	150000	12	1	2	2	
3	Active	1	25	1	2	2	1	50	450000	48	3	2	700000	12	1	2	2	
4	Active	2	38	1	12	6	3	50	190000	54	3	2	150000	12	1	2	2	
5	Active	2	30	2	6	4	3	50	60000	78	3	2	5000	12	2	2	2	
6	Active	1	40	3	11	8	3	50	500000	72	3	1	10000	12	2	2	2	
7	Active	1	43	2	20	10	3	100	300000	54	3	2	25000	12	1	2	2	
8	Active	1	36	2	8	8	1	100	25000	12	3	2	13000	12	1	2	2	
9	Active	2	47	2	12	12	4	100	2000000	12	3	1	100000	12	1	2	2	
10	Active	1	42	2	6	10	1	100	300000	48	3	3	20000	12	1	2	2	
11	Active	2	40	2	2	3	1	100	150000	48	3	2	8000	12	2	2	2	

Fig. 5.9: The data set snap shot

5.3.1.1 Preparing and exploring the data

```
> MFID = read.csv(file.choose() , sep = ',')  
> View(MFID)
```

We found that the data is structured with 18 variables and 37380 observations. The following figures explore the heading, summary and structure of the data set.

```
> head(MFID)
```

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	Active	2	32	1	8	1	1	50	1200000	56	3	2	100000	12	1	2	2	1
2	Active	2	50	2	9	9	1	50	500000	84	3	4	150000	12	1	2	2	1
3	Active	1	25	1	2	2	1	50	450000	48	3	2	700000	12	1	2	2	1
4	Active	2	38	1	12	6	3	50	190000	54	3	2	150000	12	1	2	2	1
5	Active	2	30	2	6	4	3	50	60000	78	3	2	5000	12	2	2	2	1
6	Active	1	40	3	11	8	3	50	500000	72	3	1	10000	12	2	2	2	1

```

> summary(MFID)
      AA          AB          AC          AD          AE          AF
Active :19880   Min.   :1.000   Min.   :20.00   Min.   :1.00   Min.   :1.000   Min.   :0.000
defaulter:17500 1st Qu.:2.000   1st Qu.:28.00   1st Qu.:1.00   1st Qu.:3.000   1st Qu.:1.000
              Median :2.000   Median :33.00   Median :2.00   Median :5.000   Median :3.000
              Mean   :1.929   Mean   :34.46   Mean   :1.67   Mean   :6.094   Mean   :3.704
              3rd Qu.:2.000   3rd Qu.:38.00   3rd Qu.:2.00   3rd Qu.:8.000   3rd Qu.:5.000
              Max.   :2.000   Max.   :64.00   Max.   :4.00   Max.   :30.000   Max.   :20.000

      AG          AH          AI          AJ          AK          AL
Min.   :1.000   Min.   :50.00   Min.   :2000   Min.   :12.00   Min.   :3.000   Min.   :1.000
1st Qu.:2.000   1st Qu.:50.00   1st Qu.:33000 1st Qu.:48.00   1st Qu.:3.000   1st Qu.:1.000
Median :2.000   Median :50.00   Median :72000  Median :48.00   Median :5.000   Median :2.000
Mean   :2.255   Mean   :67.23   Mean   :306090 Mean   :48.25   Mean   :4.251   Mean   :1.921
3rd Qu.:3.000   3rd Qu.:100.00 3rd Qu.:240000 3rd Qu.:48.00   3rd Qu.:5.000   3rd Qu.:3.000
Max.   :4.000   Max.   :100.00  Max.   :6900000 Max.   :105.00  Max.   :5.000   Max.   :4.000

      AM          AN          AO          AP          AQ          AR
Min.   :3000    Min.   :12.00   Min.   :1.000   Min.   :2      Min.   :1.000   Min.   :1.000
1st Qu.:10000  1st Qu.:12.00  1st Qu.:2.000  1st Qu.:2    1st Qu.:1.000  1st Qu.:1.000
Median :40000  Median :12.00  Median :2.000  Median :2    Median :1.000  Median :1.000
Mean   :175885 Mean :12.94    Mean :1.772    Mean :2      Mean :1.667    Mean :1.296
3rd Qu.:200000 3rd Qu.:12.00 3rd Qu.:2.000 3rd Qu.:2    3rd Qu.:3.000 3rd Qu.:2.000
Max.   :1000000 Max. :24.00    Max. :2.000    Max. :2      Max. :3.000    Max. :2.000

> str(MFID)
'data.frame': 37380 obs. of 18 variables:
 $ AA: Factor w/ 2 levels "Active","defaulter": 1 1 1 1 1 1 1 1 1 1 ...
 $ AB: int 2 2 1 2 2 1 1 1 2 1 ...
 $ AC: int 32 50 25 38 30 40 43 36 47 42 ...
 $ AD: int 1 2 1 1 2 3 2 2 2 2 ...
 $ AE: int 8 9 2 12 6 11 20 8 12 6 ...
 $ AF: int 1 9 2 6 4 8 10 8 12 10 ...
 $ AG: int 1 1 1 3 3 3 3 1 4 1 ...
 $ AH: int 50 50 50 50 50 50 100 100 100 100 ...
 $ AI: int 1200000 500000 450000 190000 60000 500000 300000 25000 2000000 300000 ...
 $ AJ: int 56 84 48 54 78 72 54 12 12 48 ...
 $ AK: int 3 3 3 3 3 3 3 3 3 3 ...
 $ AL: int 2 4 2 2 2 1 2 2 1 3 ...
 $ AM: int 100000 150000 700000 150000 5000 10000 25000 13000 100000 20000 ...
 $ AN: int 12 12 12 12 12 12 12 12 12 12 ...
 $ AO: int 1 1 1 1 2 2 1 1 1 1 ...
 $ AP: int 2 2 2 2 2 2 2 2 2 2 ...
 $ AQ: int 2 2 2 2 2 2 2 2 2 2 ...
 $ AR: int 1 1 1 1 1 1 1 1 1 1 ...

```

Fig. 5.10: The data structure and summary of the data set snap shot

5.3.1.1.1 Normalizing numeric data

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information [60]. This feature is important since the scale used for the values for each variable might be different. The best practice is to normalize the data and transform all the values to a common scale.

```

> normalize <- function(x) { return ((x - min(x)) / (max(x) - min(x))) }
> MFID_n <- as.data.frame(lapply(MFID[2:18], normalize))

```

After run the above code in R studio, it normalized all numeric features in the data set, instead of normalizing each of the 18 individual attributes.

The first attributes in the data set is “AA” (Loan_status) which is not numeric in nature. So, normalizing start from 2nd attributes. The function lapply () applies to normalize () each feature in the data frame. The final result is stored to MFID_n data frame using as.data.frame() function.

```
> normalize <- function(x) { return ((x - min(x)) / (max(x) - min(x))) }
> MFID_n <- as.data.frame(lapply(MFID[2:18], normalize))

> summary(MFID_n)
```

AB		AC		AD		AE		AF		AG	
Min.	:0.0000	Min.	:0.0000	Min.	:0.0000	Min.	:0.0000	Min.	:0.00000	Min.	:0.0000
1st Qu.	:1.0000	1st Qu.	:0.1915	1st Qu.	:0.0000	1st Qu.	:0.0250	1st Qu.	:0.03125	1st Qu.	:0.3333
Median	:1.0000	Median	:0.2979	Median	:0.3333	Median	:0.0750	Median	:0.09375	Median	:0.6667
Mean	:0.8265	Mean	:0.3291	Mean	:0.3222	Mean	:0.1055	Mean	:0.11422	Mean	:0.4877
3rd Qu.	:1.0000	3rd Qu.	:0.4255	3rd Qu.	:0.3333	3rd Qu.	:0.1500	3rd Qu.	:0.15625	3rd Qu.	:0.6667
Max.	:1.0000	Max.	:1.0000	Max.	:1.0000	Max.	:1.0000	Max.	:1.00000	Max.	:1.0000

AH		AI		AJ		AK		AL	
Min.	:0.0000	Min.	:0.0000000	Min.	:0.0000	Min.	:0.0000	Min.	:0.00000
1st Qu.	:0.4444	1st Qu.	:0.0000092	1st Qu.	:0.2308	1st Qu.	:0.2500	1st Qu.	:0.00000
Median	:1.0000	Median	:0.0000222	Median	:0.2308	Median	:0.7500	Median	:0.00000
Mean	:0.8043	Mean	:0.0002806	Mean	:0.2846	Mean	:0.6516	Mean	:0.09642
3rd Qu.	:1.0000	3rd Qu.	:0.0000546	3rd Qu.	:0.3077	3rd Qu.	:1.0000	3rd Qu.	:0.10000
Max.	:1.0000	Max.	:1.0000000	Max.	:1.0000	Max.	:1.0000	Max.	:1.00000

AM		AN		AO		AP		AQ		AR	
Min.	:0.00000	Min.	:0.0000	Min.	:0.0000	Min.	:0.0000	Min.	:0.0000	Min.	:0.0000
1st Qu.	:0.00070	1st Qu.	:0.3158	1st Qu.	:0.0000	1st Qu.	:0.5000	1st Qu.	:0.0000	1st Qu.	:0.0000
Median	:0.00235	Median	:0.3158	Median	:1.0000	Median	:0.5000	Median	:1.0000	Median	:1.0000
Mean	:0.01037	Mean	:0.5291	Mean	:0.6562	Mean	:0.4774	Mean	:0.5678	Mean	:0.5511
3rd Qu.	:0.00595	3rd Qu.	:0.9474	3rd Qu.	:1.0000	3rd Qu.	:0.5000	3rd Qu.	:1.0000	3rd Qu.	:1.0000
Max.	:1.00000	Max.	:1.0000	Max.	:1.0000	Max.	:1.0000	Max.	:1.0000	Max.	:1.0000

Fig 5.11 :Result of nrmalized data snap shot

5.3.1.1.2 Creating Training and Test Data Set

The kNN algorithm is applied to the training data set and the results are verified on the test data set. In this study, we would divide the data set into two portions in the ratio of 80: 20 for the training and test data set respectively. Then MFID_n data frame divided into MFID_train and MFID_test data frames. Our target variable is “AA” (loan status) which is not included in the training and test data sets.


```

> MFID_train <-MFID_n[1:29400,]
> MFID_test <- MFID_n[29401:37380,]
> MFID_train_labels <- MFID_n[1:29400,1]
> MFID_test_labels <- MFID_n[29401:37380,1]

```

5.3.1.2 Training a Model on Data Set

The `knn()` function needs to be used to train a model for which we need to install a package 'class'. The `knn()` function identifies the k-nearest neighbors using Euclidean distance where k is a user-specified number. The following command type to use `knn()`.

```

> install.packages("class")
Installing package into 'C:/Users/hp/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/class_7.3-15.zip'
Content type 'application/zip' length 106670 bytes (104 KB)
downloaded 104 KB

package 'class' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\hp\AppData\Local\Temp\Rtmp0Ye1qt\downloaded_packages
> library(class)

```

After `library(class)` is downloaded and installed in R, it is ready to use the `knn()` function to classify test data. KNN Algorithm is based on feature similarity, Choosing the right value of k is important for better accuracy. As suggested in [61], the best way to choose the value of K are the square root of the average number of complete cases (the number of observations) ($k=\sqrt{n}$). The reason they suggest this model is that k should be large enough to give a reliable result. The experiment one conducted by using 193 as a value of K, that is around the square root of the observations 37380.

`knn()` returns a factor value of predicted labels for each of the examples in the test data set which is then assigned to the data frame `MFID_test_pred`.

```

> MFID_train <-MFID_n[1:29400,]
> MFID_test <- MFID_n[29401:37380,]
> MFID_train_labels <- MFID_n[1:29400,1]
> MFID_test_labels <- MFID_n[29401:37380,1]
> MFID_test_predE1 <- knn(train = MFID_train,test = MFID_test,c1 = MFID_train_labels, k=193)

```

5.3.1.3 Evaluate the Model Performance

The model performance was evaluated using Accuracy and Precision, the model was built through KNN algorithm but we also need to check the accuracy and precision of the predicted values in MFID_test_pred as to whether they match up with the known values in MFID_test_labels. To ensure this, we need to use the CrossTable() function available in the package 'gmodels'.

```
> CrossTable(x= MFID_test_labels,y=MFID_test_predE1,prop.chisq=FALSE)
```

```
Cell contents
-----
      N
N / Row Total
N / Col Total
N / Table Total
-----
```

```
Total Observations in Table: 7980
```

MFID_test_labels	MFID_test_predE1		Row Total
	0	1	
0	555 0.988 1.000 0.070	7 0.012 0.001 0.001	562 0.070
1	0 0.000 0.000 0.000	7418 1.000 0.999 0.930	7418 0.930
Column Total	555 0.070	7425 0.930	7980

```
> ta <- table( MFID_test_predE1,MFID_test_labels)
> ta
      MFID_test_labels
MFID_test_predE1  0  1
0      555  0
1       7 7418
```

Fig. 5.12: the cross table result snap shot of experiment one

The test data consisted of 7980 observations. Out of which 7980 cases, have been accurately predicted (TP->True Positive) as Active in nature which constitutes 7418. Also, 7 out of 37380

observations were predicted (FN-> False Negative) as Active in nature but got predicted as Defaulter which constitutes 0.01%.

There is no False positive prediction that is defaulter in nature and predicted as Active. Also, 555 out of 37380 observations were predicted as not active (defaulter) correctly identified not active (defaulter) in nature is (TN -> True Negative).

The total Accuracy of the model is 99.91%, It is defined as the ratio of correctly classified instances to total number of instances [22].

```
> ta <- table( MFID_test_predE1,MFID_test_labels)
> ta
      MFID_test_labels
MFID_test_predE1  0    1
0      555    0
1      7 7418
> MFID_test_predE1 <- 100 * sum(MFID_test_labels == MFID_test_predE1)/NROW(MFID_test_labels)
> MFID_test_predE1
[1] 99.91228
```

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \dots\dots\dots (4.1)$$

$$\text{Accuracy} = 7418 + 555 / 7980 = 0.9991 = 99.91\%$$

Precision is the ratio of modules correctly classified to the number of entire modules classified fault-prone. It is proportion of units correctly predicted as faulty [22]. Precision of the model was evaluated using the following equation.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \dots\dots\dots (4.2)$$

$$\text{Precision} = 7418 / 7418 + 0 = 1$$

5.3.1.4 Improve the Performance of the Model

This can be taken into account by repeating the steps, training a model and Evaluation of model performance by changing the k-value. Generally, the K value is the square root of the observations and in this case, we took k=192 and 194 which is around a square root of 37380(193). The k-value may be fluctuated in and around the value of 193 to check the increased accuracy of the model and to keep the value of FN's as low as possible.

5.3.2 Experiment Two

5.3.2.1. Training a Model on Data Set

The experiment two conducted only by changing the value of K to 192 that is around the square root of the observations 37380 (193).

```
> MFID_train <-MFID_n[1:29400,]
> MFID_test <- MFID_n[29401:37380,]
> MFID_train_labels <- MFID_n[1:29400,1]
> MFID_test_labels <- MFID_n[29401:37380,1]

> MFID_test_predE2 <- knn(train = MFID_train,test = MFID_test,c1 = MFID_train_labels, k=192)
> CrossTable(x= MFID_test_labels,y=MFID_test_predE2,prop.chisq=FALSE)
```

```
cell contents
-----|
|                N |
| N / Row Total |
| N / Col Total  |
| N / Table Total|
|-----|
```

Total observations in Table: 7980

MFID_test_labels	MFID_test_predE2		Row Total
	0	1	
0	552 0.982 1.000 0.069	10 0.018 0.001 0.001	562 0.070
1	0 0.000 0.000 0.000	7418 1.000 0.999 0.930	7418 0.930
Column Total	552 0.069	7428 0.931	7980

Fig. 5.13: the cross table result snap shot of experiment two

5.3.2.2. Evaluate the Model Performance

The test data consisted of 7980 observations. Out of which 7980 cases have been accurately predicted (TP->True Positive) as Active in nature which constitutes 7418. Also, 10 out of 7980 observations were predicted (FN-> False Negative) as Active in nature but got predicted as Defaulter which constitutes 0.01%.

There is no False positive prediction again that is defaulter in nature and predicted as Active. Also, 552 out of 7980 observations were predicted as not active (defaulter) correctly identified not active (defaulter) in nature is (TN -> True Negative).

```
> ta <- table( MFID_test_predE2,MFID_test_labels)
> ta
      MFID_test_labels
MFID_test_predE2  0    1
0      552    0
1      10 7418
> MFID_test_predE2 <- 100 * sum(MFID_test_labels == MFID_test_predE2)/NROW(MFID_test_labels)
> MFID_test_predE2
[1] 99.87469
```

$$\text{Accuracy} = 7418 + 552/7980 = 0.9987 = 99.87\%$$

precision of the model evaluates through the following equation.

$$\text{Precision} = 7418/7418+0 = 1$$

5.3.3 Experiment Three

In experiment one and two, the accuracy and precision of the model is almost close to equal, in this experiment three perform by changing the value of K to 194 that is around the square root of the observations 37380 (193).

5.3.3.1 Training a Model

```
> MFID_test_predE3 <- knn(train = MFID_train,test = MFID_test,c1 = MFID_train_labels, k=194)
> CrossTable(x= MFID_test_labels,y=MFID_test_predE3,prop.chisq=FALSE)
```

```
Cell Contents
```

	MFID_test_predE3		Row Total
MFID_test_labels	0	1	
0	552 0.982 1.000 0.069	10 0.018 0.001 0.001	562 0.070
1	0 0.000 0.000 0.000	7418 1.000 0.999 0.930	7418 0.930
Column Total	552 0.069	7428 0.931	7980

Total observations in Table: 7980

Fig. 5.14: the cross table result snap shot of experiment three

5.3.1.2 Confusion Matrix Table

```
> ta <- table( MFID_test_predE3,MFID_test_labels)
> ta
      MFID_test_labels
MFID_test_predE3  0  1
0      552  10
1      10 7418
> MFID_test_predE3 <- 100 * sum(MFID_test_labels == MFID_test_predE3)/NROW(MFID_test_labels)
> MFID_test_predE3
[1] 99.87469
```

5.3.3.3 Evaluate the Model Performance

The test data consisted of 7980 observations. Out of which 7980 cases have been accurately predicted (TP->True Positive) as Active in nature which constitutes 7418. Also, 10 out of 7980 observations were predicted (FN-> False Negative) as Active in nature but got predicted as Defaulter which constitutes 0.01%.

There is no False positive prediction again in experiment three, that is defaulter in nature and predicted as Active. Also,552 out of 7980 observations were predicted as not active (defaulter) correctly identified not active (defaulter) in nature is (TN -> True Negative).

5.3.3.3.1 Calculate Accuracy and Precision

$$\text{Accuracy} = 7418 + 552/7980 = 0.9987 = 99.87\%$$

precision of the experiment three model evaluates through the following equation.

$$\text{Precision} = 7418/7418+0= 1$$

5.3.4 Experiment Four

In this experiment four we test different k values to check improvement performance of the model. The following below R code shows accuracy level for different K values (185-205).

```
> i=1
>
> k.optm=1
> for (i in 185:205){
+     knn.mod <- knn(train=MFID_train, test=MFID_test, cl=MFID_train_labels, k=i)
+     k.optm[i] <- 100 * sum(MFID_test_labels == knn.mod)/NROW(MFID_test_labels)
+     k=i
+     cat(k, '=', k.optm[i], '\n') }
185 = 99.87469
186 = 99.88722
187 = 99.83709
188 = 99.83709
189 = 99.83709
190 = 99.82456
191 = 99.84962
192 = 99.86216
193 = 99.88722
194 = 99.83709
195 = 99.86216
196 = 99.88722
197 = 99.7995
198 = 99.81203
199 = 99.77444
200 = 99.91228
201 = 99.84962
202 = 99.84962
203 = 99.88722
204 = 99.86216
205 = 99.88722
```

Fig. 5.15: the training model of different K values and result snap shot

```
> plot(k.optm, type="b", xlab="K- value",ylab="Accuracy level")
```

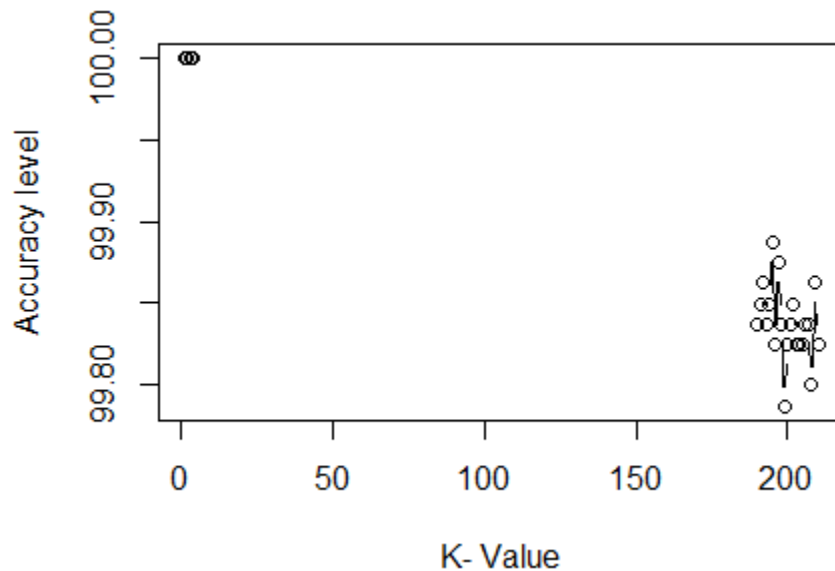


Fig. 5.16: The training model of different K values and result plot

5.3.5 Attribute(variable) Importance

In this section, to identify attribute importance we use the earth package in R to test variable importance based on Generalized cross validation (GCV), number of subsets models the variable occurs (nsubsets) and residual sum of squares (RSS).

```
> knnmodelimportance <- earth(AA ~ ., data=MFID)
> ev <- evimp (knnmodelimportance)
> ev
  nsubsets  gcv  rss
AQ      26 100.0 100.0
AG      25  53.8  53.9
AM      23  35.0  35.1
AO      20  27.1  27.2
AL      19  24.2  24.3
AI      18  22.0  22.1
AC      15  16.7  16.9
AE      14  15.5  15.6
AF      12  13.7  13.8
AD      10  10.1  10.3
AJ       8   7.2   7.4
> plot(ev)
```

Fig. 5.17: attribute importance snap shot

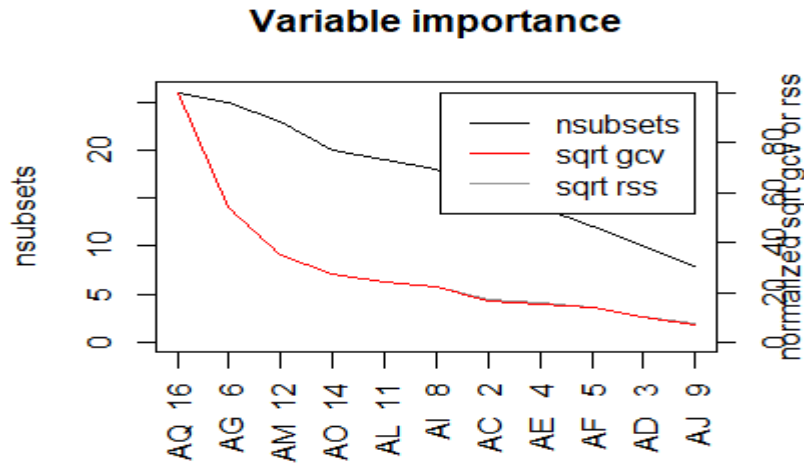


Fig. 5.18: Attribute importance plot

As stated in the above figure, eleven attributes are listed as importance variables out of 17 independent attributes, but the two attributes contributed massively in the prediction are, AQ (MFIs organization Type attributes) and AG (business sector attributes).

5.4. Logistic Regression Modeling

In this experiment a Prediction model building is done using Logistic regression algorithm. Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No [35]. The main thing hear is to determine a mathematical equation that can be used to predict the probability of event 1(active). After the equation is established, it can be used to predict the Y (loan status) attributes when only the X's (other 17 attributes) are known using selected microfinance data set through "mlbench" package.

5.4.1. Experiment one

5.4.1.1 Install additional needed Package in to R studio

Install "mlbenh" package and load the data in to R studio and keep only the complete cases.

```

> install.packages("mlbench")
Installing package into 'C:/Users/hp/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/mlbench_2.1-1.zip'
Content type 'application/zip' length 1059296 bytes (1.0 MB)
downloaded 1.0 MB

package 'mlbench' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\hp\AppData\Local\Temp\RtmpCK6Bu8\downloaded_packages
> LG = read.csv(file.choose() , sep = ',')
> View(LG)

```

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	1	2	32	1	8	1	1	50	1200000	56	3	2	100000	12	1	2	2	1
2	1	2	50	2	9	9	1	50	500000	84	3	4	150000	12	1	2	2	1
3	1	1	25	1	2	2	1	50	450000	48	3	2	700000	12	1	2	2	1
4	1	2	38	1	12	6	3	50	190000	54	3	2	150000	12	1	2	2	1
5	1	2	30	2	6	4	3	50	60000	78	3	2	5000	12	2	2	2	1
6	1	1	40	3	11	8	3	50	500000	72	3	1	10000	12	2	2	2	1
7	1	1	43	2	20	10	3	100	300000	54	3	2	25000	12	1	2	2	1
8	1	1	36	2	8	8	1	100	25000	12	3	2	13000	12	1	2	2	1
9	1	2	47	2	12	12	4	100	2000000	12	3	1	100000	12	1	2	2	1
10	1	1	42	2	6	10	1	100	300000	48	3	3	20000	12	1	2	2	1
11	1	2	40	2	2	3	1	100	150000	48	3	2	8000	12	2	2	2	1

Fig. 5.19: The data set view snapshot for LR model

The dataset has 37,380 observations and 18 attributes. The AA (loan status) column is the response (dependent) attributes and it tells is it active or defaulter. Structure of the data set shown in the following figure: -

```

> str(LG)
'data.frame': 37380 obs. of 18 variables:
 $ AA: int 1 1 1 1 1 1 1 1 1 1 ...
 $ AB: int 2 2 1 2 2 1 1 1 2 1 ...
 $ AC: int 32 50 25 38 30 40 43 36 47 42 ...
 $ AD: int 1 2 1 1 2 3 2 2 2 2 ...
 $ AE: int 8 9 2 12 6 11 20 8 12 6 ...
 $ AF: int 1 9 2 6 4 8 10 8 12 10 ...
 $ AG: int 1 1 1 3 3 3 3 1 4 1 ...
 $ AH: int 50 50 50 50 50 50 100 100 100 ...
 $ AI: int 1200000 500000 450000 190000 60000 500000 300000 25000 2000000 300000 ...
 $ AJ: int 56 84 48 54 78 72 54 12 12 48 ...
 $ AK: int 3 3 3 3 3 3 3 3 3 3 ...
 $ AL: int 2 4 2 2 2 1 2 2 1 3 ...
 $ AM: int 100000 150000 700000 150000 5000 10000 25000 13000 100000 20000 ...
 $ AN: int 12 12 12 12 12 12 12 12 12 12 ...
 $ AO: int 1 1 1 1 2 2 1 1 1 1 ...
 $ AP: int 2 2 2 2 2 2 2 2 2 2 ...
 $ AQ: int 2 2 2 2 2 2 2 2 2 2 ...
 $ AR: int 1 1 1 1 1 1 1 1 1 1 ...

```

Fig. 5.20: the structure of the data set view snapshot

The following code in R enable to convert the response variable AA to a factor variable and all other columns to numeric.

```
> for(i in 1:18) {  
+   LG[, i] <- as.numeric(as.character(LG[, i]))  
+ }  
  
> LG$AA <- ifelse(LG$AA == "active", 1, 0)  
> LG$AA <- factor(LG$AA, levels = c(0, 1))  
> |
```

5.4.1.2 Balance the Imbalance Data Set

It is known that the data was split into 80:20 training and test samples ratio. As the response attributes (loan status) is a binary categorical variable, there is a need to balance the training data into equal proportion of classes.

```
> table(LG2M$AA)  
  
 0    1  
17500 19880  
> |
```

The loan status active and defaulter are split approximately in 1:2 ratios. As shows in above code result, clearly there is a few loan status class imbalances. So, before building the logistic regression model, it needs to build the samples such that both the 1's and 0's (active and defaulter) are in approximately equal proportions. This concern is normally handled with a couple of techniques called: Down sampling, Up sampling and Hybrid Sampling using SMOTE and ROSE[57].

In Down sampling, the majority class is randomly down sampled to be of the same size as the smaller class. That means, when creating the training dataset, the rows with the defaulter Class will be picked fewer times during the random sampling.

Similarly, in Up Sampling, rows from the minority class, that is, defaulter is repeatedly sampled over and over till it reaches the same size as the majority class (active). But in case of Hybrid sampling (SMOTE and ROSE packages), artificial data points are

generated and are systematically added around the minority class [57]. In this study we use up sampling techniques to balance the imbalance data. So first we create the Training and Test data using “caret” Package in RStudio.

```
> library(caret)
Loading required package: lattice
Loading required package: ggplot2
Warning messages:
1: package 'caret' was built under R version 3.5.3
2: package 'ggplot2' was built under R version 3.5.3

> '%ni%' <- Negate('%in%')
> options(scipen=999)
> set.seed(100)

> trainDataIndex <- createDataPartition(LG2M$AA, p=0.7, list = F)
> trainData <- LG2M[trainDataIndex, ]
> trainDataIndex <- createDataPartition(LG2M$AA, p=0.8, list = F)
> trainData <- LG2M[trainDataIndex, ]

> testData <- LG2M[-trainDataIndex, ]
```

In the above R studio code snapshot, shows installed caret package and used the “createDataPartition” function to generate the row numbers for the training dataset. As stated in the above code p=0.8 that is 80% of the data set rows to go inside “trainData” for training the model and the remaining 20% to go to “testData” for test.

```
> table(trainData$AA)

  0    1
14000 15904
```

There is around 2000 rows more active samples than defaulter sample. So, it needs up sampling to balance the data set using the “upSample” function in R.

```
> up_train <- upSample(x = trainData[, colnames(trainData) %ni% "class"],y = trainData$AA)
> table(up_train$AA)

  0    1
15904 15904
```

As a result of the above up sampling code result, active and defaulter status are now in the same ratio.

5.4.1.3 Building the Logistic Regression Model with Balanced Data Set

```
> logitmod <- glm(Class ~ AB + AC + AD + AE + AF + AG + AH + AI + AJ + AK + AL + AM + AN + AO + AP + AQ, family = "binomial", data=up_train)

> summary(logitmod)
```

```
Call:
glm(formula = Class ~ AB + AC + AD + AE + AF + AG + AH + AI +
    AJ + AK + AL + AM + AN + AO + AP + AQ, family = "binomial",
    data = up_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8042  -0.3050  -0.0001   0.0000   4.0663

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -94.91053053182 2702.81462644313  -0.035    0.972
AB           -6.88696458289   824.07346749885  -0.008    0.993
AC            0.04802698126    0.00295906797  16.230 < 0.0000000000000002 ***
AD           -0.57180168213    0.03643342525 -15.694 < 0.0000000000000002 ***
AE           -0.00773728134    0.00607321169  -1.274    0.203
AF            0.00762890982    0.00796267773   0.958    0.338
AG           -3.02312579741    0.05022210926 -60.195 < 0.0000000000000002 ***
AH           -0.20592820055    15.80742622309  -0.013    0.990
AI           -0.00000090236    0.0000006324 -14.269 < 0.0000000000000002 ***
AJ           -0.26754128577    11.55715782353  -0.023    0.982
AK           14.55713956097   197.63434005406   0.074    0.941
AL            1.43102725658    0.03679378270  38.893 < 0.0000000000000002 ***
AM            0.00000090613    0.00000007882  11.496 < 0.0000000000000002 ***
AN           -0.04451633605    0.00757400866  -5.878    0.00000000416 ***
AO            1.58051134743    0.05286984448  29.894 < 0.0000000000000002 ***
AP                NA                NA                NA                NA
AQ            58.20881132411   586.46205101590   0.099    0.921
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44095  on 31807  degrees of freedom
Residual deviance: 13239  on 31792  degrees of freedom
AIC: 13271

Number of Fisher Scoring iterations: 21
```

Fig. 5.21: The result of logistic regression model snap shot

5.4.1.4 Interpretation of The Logistic Model Result

After building the logistic regression model, then we analyze the fitting and interpret what the model is telling us. In above R code logistic regression model result, the “glm” function internally encodes categorical variables into $n - 1$ distinct level. The column “Estimate” represents the regression coefficients value. Here, the regression coefficients explain the change in $\log(\text{odds})$ (The odds are itself the ratio of two probabilities, p and $1-p$) of the response variable for one-unit change in the predictor variable.

The column “Std. Error” in above logistic model result represents the standard error associated with the regression coefficients. And z value is analogous to t -statistics in multiple regression output. z value > 2 implies the corresponding attributes is significant.

p value determines the probability of significance of predictor variables. A variable having $p < 0.5$ is considered an important predictor [35]. A predictor that has a low p -value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable [35].

In the existence of other attributes, attributes like AB (sex), AE (total years of any business experience), AF(years since current business experience), AH(percentage of collateral ownership), AJ(hours worked per week), and AK(loan product type) are statically not significant ($p > 0.5$). As for the statistically significant variables, AL (loan cycle), AM(approved loan amount), AN(loan term in month), AO(outstanding principal balance), AC(age), AG(business sector), AI(business yearly earning) and AD(borrower's education status) has the lowest p -value that indicate a strong association of those attributes to the loan with the probability of having active borrowers. AIC (Akaike information criterion) value of this model is 13271, then we create another model to achieve a lower AIC value without including such not significant attributes.

```
> logitmoda <- glm(AA ~ AC + AD + AG + AI + AL + AM + AN + AO + AP, family = "binomial", data=up_train)
> summary(logitmod)
```

```

Call:
glm(formula = AA ~ AB + AC + AD + AE + AF + AG + AH + AI + AJ +
     AK + AL + AM + AN + AO + AP + AQ, family = "binomial", data = trainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8586  -0.2598   0.0000   0.0000   3.9961

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -94.95829583336 2642.24116181527  -0.036  0.9713
AB           -6.89071667007  812.16964124293  -0.008  0.9932
AC            0.04832978851   0.00301925439  16.007 < 0.0000000000000002 ***
AD           -0.55762011630   0.03699594601 -15.072 < 0.0000000000000002 ***
AE           -0.01162722569   0.00621323541  -1.871  0.0613 .
AF            0.01224616194   0.00814044110   1.504  0.1325
AG           -2.96512456802   0.05054931674 -58.658 < 0.0000000000000002 ***
AH           -0.20565061575   15.04343863820  -0.014  0.9891
AI           -0.00000092002   0.00000006451 -14.263 < 0.0000000000000002 ***
AJ           -0.26588482521   10.97081292092  -0.024  0.9807
AK           14.60123714320   199.74234351835   0.073  0.9417
AL            1.39528743203   0.03731341973  37.394 < 0.0000000000000002 ***
AM            0.00000094978   0.00000008006  11.863 < 0.0000000000000002 ***
AN           -0.04546025818   0.00774100300  -5.873  0.00000000429 ***
AO            1.52976745547   0.05381202877  28.428 < 0.0000000000000002 ***
AP              NA              NA              NA              NA
AQ           58.09113051961   579.65149033314   0.100  0.9202
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41334  on 29903  degrees of freedom
Residual deviance: 12580  on 29888  degrees of freedom
AIC: 12612

```

Fig. 5.22: The result of logistic regression model with significant attributes snap shot.

As a result of the above model (with excluding not significant attribute) achieved with a lower AIC value (12612) and a better model.

5.4.1.5 Predict on Test Dataset

To predict the observation using logistic regression we need to set type="response" in order to compute the prediction probabilities.

```
> pred <- predict(logitmod, newdata = testData, type = "response")
```

The prediction "pred" contains the probability that the observation is active or defaulter for each observation. The common practice is to take the probability cutoff as 0.5. If the probability of Y is > 0.5, then it can be classified an event (Active). So, if "pred" is greater than 0.5, it is Active else it is Defaulter. The following R code shows the prediction of the observation.

```
> pred <- predict(logitmod, newdata = testData, type = "response")
```

```

> y_pred_num <- ifelse(pred > 0.5, 1, 0)
> y_pred <- factor(y_pred_num, levels=c(0, 1))
> y_act <- testData$AA
> mean(y_pred == y_act)

```

5.4.1.6 Evaluate Model Performance

Then after prediction we compute the accuracy using proportion of `y_pred` that matches with `y_act`.

```

> mean(y_pred == y_act)
[1] 0.9242911

```

So, the total accuracy rate of the model using proportion of `y_pred` with `y_act` is 92.4% as shows in the above R code result of the accuracy.

```

> CrossTable(x = y_pred, y = y_act, prob = FALSE)

```

```

Cell Contents
-----
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total

```

```

Total Observations in Table: 7476

```

y_pred	y_act		Row Total
	0	1	
0	3327 1443.281 0.894 0.951 0.445	393 1270.494 0.106 0.099 0.053	3720 0.498
1	173 1429.447 0.046 0.049 0.023	3583 1258.316 0.954 0.901 0.479	3756 0.502
Column Total	3500 0.468	3976 0.532	7476

Fig. 5.23: Cross table result snap shot

To create the confusion matrix table and calculate the accuracy of the model, the package “e1071” is installed and run the confusion matrix function like below code in RStudio.

```
install.packages('e1071', dependencies=TRUE)
```

```
> confusionMatrix(y_pred, y_act)
Confusion Matrix and Statistics

      Reference
Prediction  0      1
 0  3327  393
 1   173 3583

      Accuracy : 0.9243
      95% CI   : (0.9181, 0.9302)
  No Information Rate : 0.5318
  P-Value [Acc > NIR] : < 0.000000000000000022

      Kappa : 0.8485

  Mcnemar's Test P-Value : < 0.000000000000000022

      Sensitivity : 0.9506
      Specificity : 0.9012
  Pos Pred Value : 0.8944
  Neg Pred Value : 0.9539
    Prevalence   : 0.4682
  Detection Rate : 0.4450
  Detection Prevalence : 0.4976
  Balanced Accuracy : 0.9259

      'Positive' Class : 0
```

Fig. 5.24: confusion matrix results snap shot

Precision of the model evaluates through the following equation. $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$

$$\text{Precision} = 3583 / (3583 + 393) = 0.901$$

To plot the results of sensitivity and specificity of the model and to check what threshold achieves (calculate the AUC) we used rock curve as a tool using “ROCR” function. The ROC is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC (area under the curve) is the area under the ROC (Receiver Operator Characteristic) curve. A model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5[67].

```
> plot(y_act, y_pred)
> library(ROCR)

> ROCRpred = prediction(pred, y_act)
> ROCRperf = performance(ROCRpred, "tpr", "fpr")
> plot(ROCRperf, colorize = TRUE)
```

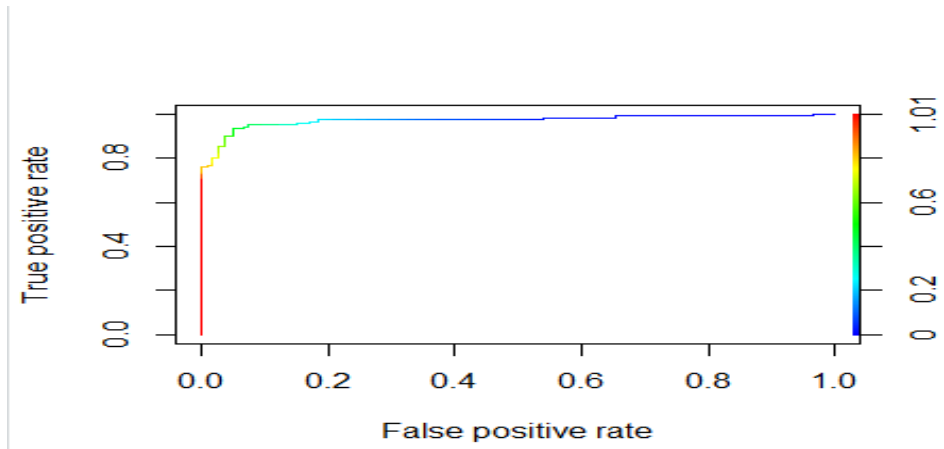


Fig. 5.25: confusion matrix results snap shot

```
> library(InformationValue)
> plotROC(y_act, pred)
```

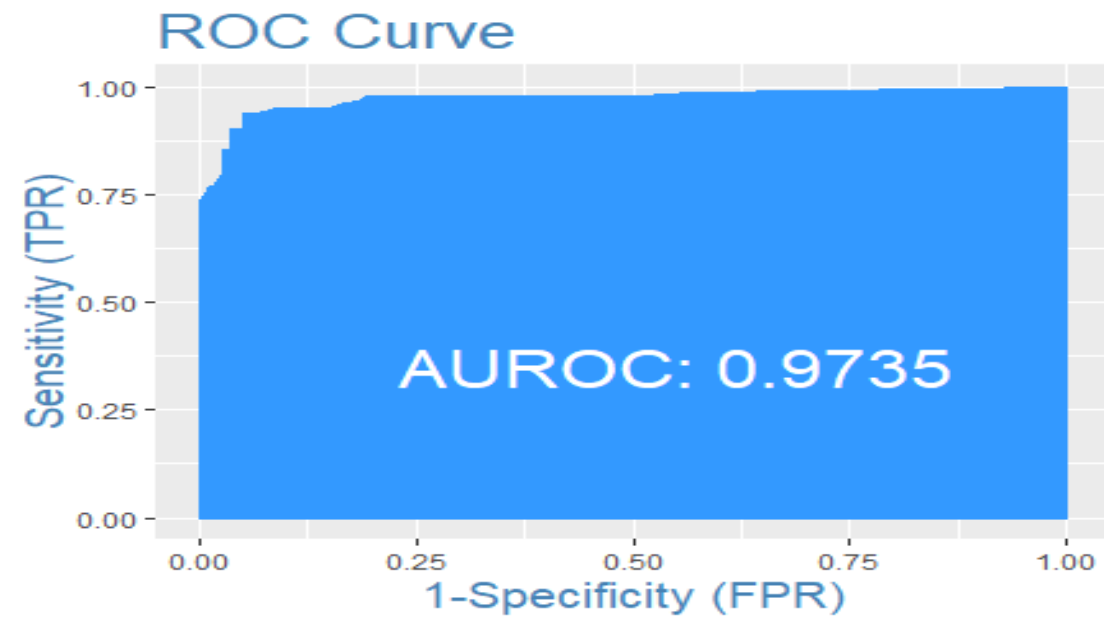


Fig. 5.26: ROC Curve result snap shot

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 index of accuracy (A) or concordant index, represents the performance of the ROC curve. A good model, the curve should rise steeply, indicating that

the TPR (Y-Axis) increases faster than the FPR (X-Axis) as the cutoff score decreases. Greater the area under the ROC curve, better the predictive ability of the model. (Higher the area, better the model) [67]. As showed in the above diagram ROC is plotted between True Positive Rate (Y axis) and False Positive Rate (X Axis). In the above plot, the area under curve cover the maximum area (ROC curve 97.35%) and the curve is higher and close to the left corner (true positive) that indicate the model is pretty good.

5.4.2. Experiment two

The experiment two conducted by using unbalanced data set unlike to the experiment one.

5.4.2.1 Building the Logistic Regression Model with Imbalance Data Set

```
> logitmod <- glm(AA ~ AB + AC + AD + AE + AF + AG + AH + AI + AJ + AK + AL + AM + AN + AO + AP + AQ, family
= "binomial", data=trainData)
```

```
> pred <- predict(logitmod, newdata = testData, type = "response")
```

```
> y_pred_num <- ifelse(pred > 0.5, 1, 0)
> y_pred <- factor(y_pred_num, levels=c(0, 1))
> y_act <- testData$AA
> mean(y_pred == y_act)
[1] 0.9388711
```

```
> CrossTable(x = y_pred,y= y_act,prob.chisq=FALSE)
```

```
Cell Contents
-----
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total
-----
```

```
Total observations in Table: 7476
```

y_pred	y_act		Row Total
	0	1	
0	3327 1584.101 0.921 0.951 0.445	284 1394.455 0.079 0.071 0.038	3611 0.483
1	173 1479.997 0.045 0.049 0.023	3692 1302.814 0.955 0.929 0.494	3865 0.517
Column Total	3500 0.468	3976 0.532	7476

```

> confusionMatrix(y_pred, y_act)
Confusion Matrix and Statistics

          Reference
Prediction 0      1
0      3327  284
1      173 3692

          Accuracy : 0.9389
          95% CI   : (0.9332, 0.9442)
    No Information Rate : 0.5318
    P-Value [Acc > NIR] : < 0.0000000000000000022

          Kappa   : 0.8775

    Mcnemar's Test P-Value : 0.0000002667

          Sensitivity : 0.9506
          Specificity : 0.9286
    Pos Pred Value   : 0.9214
    Neg Pred Value   : 0.9552
    Prevalence       : 0.4682
    Detection Rate   : 0.4450
    Detection Prevalence : 0.4830
    Balanced Accuracy : 0.9396

    'Positive' Class : 0

```

Fig. 5.27: cross table and confusion matrix results snap shot

Precision of the model with imbalanced data set evaluates through the following equation.

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{Precision} = \frac{3692}{3692+284} = 0.928$$

As shown in above prediction model result, the accuracy of the model is 93%. It shows that the unbalanced data set in this logistic regression model predict more accurately by 1% than the balanced (up balance) data set in experiment one.

5.5. Navies Bayes Model

In this experiment a Prediction model building is done using Naive Bayes Classification algorithm and RStudio used to predict the loans with R programing.

5.5.1 Data description

For this Navies Bayes model also, we use samples of about 37380 loan borrowers' details with 17 independent and one dependent attributes. The data set and its structure are as follows: -

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	Active	2	32	1	8	1	1	50	1200000	56	3	2	100000	12	1	2	2	1
2	Active	2	50	2	9	9	1	50	500000	84	3	4	150000	12	1	2	2	1
3	Active	1	25	1	2	2	1	50	450000	48	3	2	700000	12	1	2	2	1
4	Active	2	38	1	12	6	3	50	190000	54	3	2	150000	12	1	2	2	1
5	Active	2	30	2	6	4	3	50	60000	78	3	2	5000	12	2	2	2	1
6	Active	1	40	3	11	8	3	50	500000	72	3	1	10000	12	2	2	2	1
7	Active	1	43	2	20	10	3	100	300000	54	3	2	25000	12	1	2	2	1
8	Active	1	36	2	8	8	1	100	25000	12	3	2	13000	12	1	2	2	1
9	Active	2	47	2	12	12	4	100	2000000	12	3	1	100000	12	1	2	2	1
10	Active	1	42	2	6	10	1	100	300000	48	3	3	20000	12	1	2	2	1
11	Active	2	40	2	2	3	1	100	150000	48	3	2	8000	12	2	2	2	1

```

> str(NBMFIL)
'data.frame': 37380 obs. of 18 variables:
 $ AA: Factor w/ 2 levels "Active","defaulter": 1 1 1 1 1 1 1 1 1 1 ...
 $ AB: int 2 2 1 2 2 1 1 1 2 1 ...
 $ AC: int 32 50 25 38 30 40 43 36 47 42 ...
 $ AD: int 1 2 1 1 2 3 2 2 2 2 ...
 $ AE: int 8 9 2 12 6 11 20 8 12 6 ...
 $ AF: int 1 9 2 6 4 8 10 8 12 10 ...
 $ AG: int 1 1 1 3 3 3 3 1 4 1 ...
 $ AH: int 50 50 50 50 50 50 100 100 100 100 ...
 $ AI: int 1200000 500000 450000 190000 60000 500000 300000 25000 2000000 300000 ...
 $ AJ: int 56 84 48 54 78 72 54 12 12 48 ...
 $ AK: int 3 3 3 3 3 3 3 3 3 3 ...
 $ AL: int 2 4 2 2 2 1 2 2 1 3 ...
 $ AM: int 100000 150000 700000 150000 5000 10000 25000 13000 100000 20000 ...
 $ AN: int 12 12 12 12 12 12 12 12 12 12 ...
 $ AO: int 1 1 1 1 2 2 1 1 1 1 ...
 $ AP: int 2 2 2 2 2 2 2 2 2 2 ...
 $ AQ: int 2 2 2 2 2 2 2 2 2 2 ...
 $ AR: int 1 1 1 1 1 1 1 1 1 1 ...

```

Fig. 5.28: The data set and its structure for NB model snap shot

Then we do Convert the loan status field (AA) values Active and Defaulter to ‘1’and ‘0’ respectively and also Convert the loan status field as factor as shown below: -

```

> NBMFIL$AA <- ifelse(NBMFIL$AA == 'Active',1,0)
> NBMFIL$AA <- factor(NBMFIL$AA,levels = c(0,1))

```

5.5.2 Partitioning the Data Set

In predictive modeling, the data needs to be partitioned into train and test sets. 80% of the data is partitioned for training purpose and 20% of the data for testing purpose as we do in above

other models. In this section After data splitting, we apply Feature scaling to standardize the range of independent variables.

```
> set.seed(12)

> split = sample.split(NBMFIL$AA, splitRatio = 0.80)
> training_set = subset(NBMFIL, split == FALSE)
> training_set = subset(NBMFIL, split == TRUE)
> test_set = subset(NBMFIL, split == FALSE)
> training_set[-18] = scale(training_set[-18])
```

5.5.3 Classification Using Naive Bayes

In this section Naïve Bayes classification model is executed in RStudio on top of the MFIs dataset to classify Active and Defaulter borrowers. To do Naive Bayes classification model, we perform the following: - first we Install and load “e1071” package before running Naive Bayes, Test the models built using train datasets through the test dataset and then Using accuracy, precision and error rate, we analyses how these models are behaving for the test dataset.

```
> classifier = naiveBayes(x= training_set, y = training_set$AA)
> y_pred = predict(classifier, newdata = test_set)
```

The above code used to classify the dataset using Naïve Bayes and predict it using test data set. The result of the model also evaluates through accuracy and precision as we do in previous two models.

5.5.4 Visualize Test Set Result

We visualize the test set result using “crossTable” function in RStudio and summary of it using table.

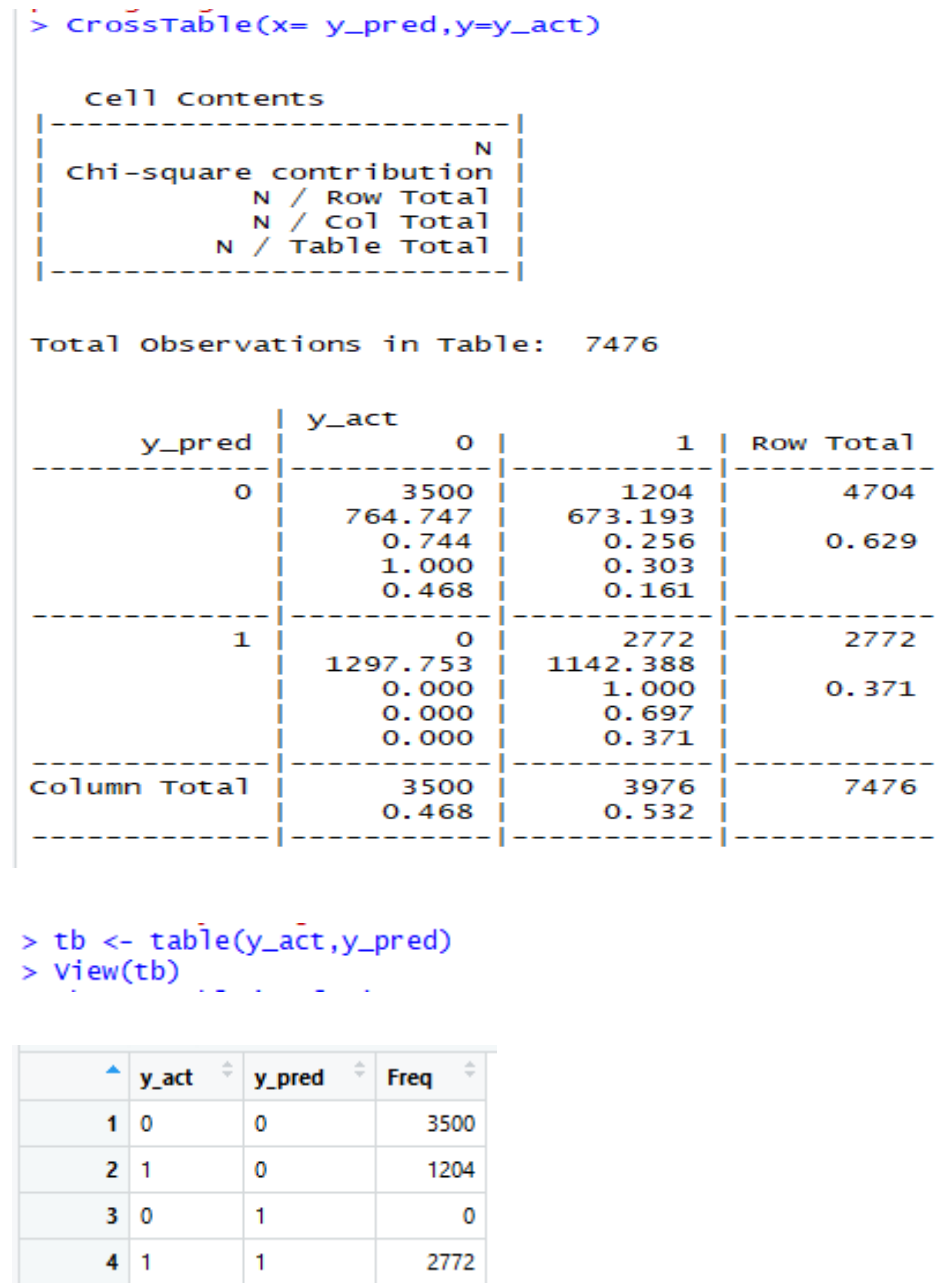


Fig. 5.29: The cross table result snap shot

5.5.5 Evaluate the Model Performance

The above Naïve Bayes model result evaluated using accuracy and precision techniques as we do in previous two models before. The test data consisted of 7476 observations. Out of which 3500 cases have been accurately predicted (TN->True Negative) as (Not active) defaulter in nature which constitutes 46.8%. Also, 1204 out of 7476 observations were predicted (FP->False Positive) as defaulter in nature but got predicted as Active which constitutes 16.1%.

Also, 2772 out of 7476 observations were correctly predicted (TP -> True Positive) as Active in nature which constitutes 37.1% and there is no (False Positive ->FP) prediction that mean defaulter in nature but got predicted as active. The total Accuracy of the model is 83.9%.

$$\text{Accuracy} = 6272 / 7476 = 83.89\%$$

precision of the model evaluates through the following equation.

$$\text{Precision} = 2772 / 2772 + 0 = 1$$

Accuracy of the model using confusion Matrix table function in RStudio.

```
> confusionMatrix(table(y_act,y_pred))
Confusion Matrix and Statistics

      y_pred
y_act  0      1
 0 3500      0
 1 1204 2772

      Accuracy : 0.839
      95% CI   : (0.8304, 0.8472)
  No Information Rate : 0.6292
  P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 0.6831

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

      Sensitivity : 0.7440
      Specificity : 1.0000
  Pos Pred Value : 1.0000
  Neg Pred Value : 0.6972
      Prevalence : 0.6292
  Detection Rate : 0.4682
  Detection Prevalence : 0.4682
  Balanced Accuracy : 0.8720

      'Positive' Class : 0
```

Fig. 5.30: confusion matrix results snap shot

5.6. Support Vector Machine (SVM) Model

SVM (Support Vector Machine) is a supervised machine learning algorithm which is mainly used to classify data into different classes. Unlike most algorithms, SVM makes use of a hyperplane which acts like a decision boundary between the various classes. These closest data points to the hyperplane are known as support vectors [30]. As we do in other previous three models, we use R programming language to build an SVM classifier model.

5.6.1 Install Caret Packages in to RStudio

The caret package is also known as the Classification and Regression Training, has tons of functions that helps to build predictive models. It contains tools for data splitting, pre-processing, feature selection, tuning, unsupervised learning algorithms [66].

```
> install.packages("caret")
```

5.6.2 Data Description and Load the Data Set

The data set we use in this model have about 37380 borrowers record details with 17 independent and one dependent attributes, it stored in a CSV or Comma Separated Version.

To see the structure of the dataset we use function str (): -

```
> str(SVMMFIL)
'data.frame':  37380 obs. of  18 variables:
 $ AA: Factor w/ 2 levels "Active","defaulter": 1 1 1 1 1 1 1 1 1 1 ...
 $ AB: int  2 2 1 2 2 1 1 1 2 1 ...
 $ AC: int  32 50 25 38 30 40 43 36 47 42 ...
 $ AD: int  1 2 1 1 2 3 2 2 2 2 ...
 $ AE: int  8 9 2 12 6 11 20 8 12 6 ...
 $ AF: int  1 9 2 6 4 8 10 8 12 10 ...
 $ AG: int  1 1 1 3 3 3 3 1 4 1 ...
 $ AH: int  50 50 50 50 50 50 100 100 100 100 ...
 $ AI: int  120000 500000 450000 190000 60000 500000 300000 25000 2000000 300000 ...
 $ AJ: int  56 84 48 54 78 72 54 12 12 48 ...
 $ AK: int  3 3 3 3 3 3 3 3 3 3 ...
 $ AL: int  2 4 2 2 2 1 2 2 1 3 ...
 $ AM: int  100000 150000 700000 150000 5000 10000 25000 13000 100000 20000 ...
 $ AN: int  12 12 12 12 12 12 12 12 12 12 ...
 $ AO: int  1 1 1 1 2 2 1 1 1 1 ...
 $ AP: int  2 2 2 2 2 2 2 2 2 2 ...
 $ AQ: int  2 2 2 2 2 2 2 2 2 2 ...
 $ AR: int  1 1 1 1 1 1 1 1 1 1 ...
```

Fig. 5.31: structure of the data set snap shot

The output shows that the dataset consists of 37380 observations each with 18 attributes.

The below code is used to convert the data frame's "AA" column to a factor variable.

```
> SVMFIL$AA <- ifelse(SVMFIL$AA == 'Active',1,0)
> SVMFIL$AA <- factor(SVMFIL$AA,levels = c(0,1))
> view(SVMFIL)
> head(SVMFIL)
```

	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	1	2	32	1	8	1	1	50	1200000	56	3	2	100000	12	1	2	2	1
2	1	2	50	2	9	9	1	50	500000	84	3	4	150000	12	1	2	2	1
3	1	1	25	1	2	2	1	50	450000	48	3	2	700000	12	1	2	2	1
4	1	2	38	1	12	6	3	50	190000	54	3	2	150000	12	1	2	2	1
5	1	2	30	2	6	4	3	50	60000	78	3	2	5000	12	2	2	2	1
6	1	1	40	3	11	8	3	50	500000	72	3	1	10000	12	2	2	2	1

5.6.3 Split the Data in to Training and Test Set

In this section the data set split into training set and testing set with 80:20 ratio, this is also called data splitting. The training set specifically used for the model building and the testing set for evaluating the model.

The caret package in R provides a method createDataPartition() which is basically for partitioning our data into train and test set.

```
> intrain <- createDataPartition(y = SVMFIL$AA, p= 0.8, list = FALSE)
> training <- SVMFIL[intrain,]
> testing <- SVMFIL[-intrain,]
```

In the above R code, the "y" parameter takes the value of variable according to which data needs to be partitioned. In our case, target variable is at AA, so we are passing SVMFIL\$AA. The "p" parameter holds a decimal value in the range of 0-1. It's to show the percentage of the split. We are using p=0.8. It means that data split should be done in 80:20 ratios. So, 80% of the data is used for training and the remaining 20% is for testing the model. The createDataPartition() method is returning a matrix "intrain". This "intrain" matrix has training data set and we're storing this in the 'training' variable and the rest of the data (20%) of it stored in the testing variable. Then the dimensions of training data frame and testing data frame is looks like the following.

```
> dim(training)
[1] 29904 18
> dim(testing)
[1] 7476 18
```

5.6.4 Train the Model

To train the model, first we need to implement the `trainControl()` method provided by the `caret` package. The training method is used to train the data on specific algorithms.

```
> trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
```

In the above code We used the “number” parameter to holds the number of resampling iterations. And the “repeats” parameter contains the sets to compute for the repeated cross-validation. We are using setting number =10 and repeats =3.

```
> svm_Linear <- train(AA ~., data = training, method = "svmLinear",trControl=trctrl,preProcess = c("center",  
"scale"),tuneLength = 10)
```

In the above train model code, the “AA~.” denotes a formula for using all attributes in the classifier and AA as the target variable. And The “preProcess” parameter is set for preprocessing the training data with passing two values parameter “center” and “scale.

We are passing 2 values in our “pre-process” parameter “center” & “scale”. These two parameters help for centering and scaling the data. After pre-processing, these convert the training data with mean value as approximately “0” and standard deviation as “1”. The “tuneLength” parameter holds an integer value. This is for tuning the algorithm. The result of the train () method is save in the `svm_Linear` variable as follows.

```
> svm_Linear  
Support Vector Machines with Linear kernel  
  
29904 samples  
  17 predictor  
   2 classes: '0', '1'  
  
Pre-processing: centered (17), scaled (17)  
Resampling: Cross-Validated (10 fold, repeated 3 times)  
Summary of sample sizes: 26914, 26914, 26913, 26913, 26913, 26914, ...  
Resampling results:  
  
Accuracy   Kappa  
0.9300427  0.859732  
  
Tuning parameter 'C' was held constant at a value of 1
```

The model tested at value “C” =1., the model is trained with C value as 1. Then we predict classes for the test set using predict () method. The caret package provides predict () method for predicting results. We are passing 2 arguments. Its first parameter is the trained model and second parameter “newdata” holds the testing data frame. The predict () method returns a list, then we save it in a test_pred variable.

```
> test_pred <- predict(svm_linear, newdata = testing)
```

5.6.5 Model Performance Evaluation

To check the performance of the model through accuracy and precision we use the confusion matrix. The result of confusion matrix is shown in below code.

```
> confusionMatrix(table(test_pred, testing$AA))
Confusion Matrix and Statistics

test_pred   0   1
  0 3246  314
  1   254 3662

      Accuracy : 0.924
      95% CI   : (0.9178, 0.9299)
 No Information Rate : 0.5318
 P-Value [Acc > NIR] : <2e-16

      Kappa   : 0.8476

 Mcnemar's Test P-Value : 0.0133

      Sensitivity : 0.9274
      Specificity : 0.9210
   Pos Pred Value : 0.9118
   Neg Pred Value : 0.9351
      Prevalence  : 0.4682
   Detection Rate : 0.4342
   Detection Prevalence : 0.4762
   Balanced Accuracy : 0.9242

   'Positive' Class : 0
```

Fig. 5.32: Confusion matrix snap shot

The output of confusion matrix shows that the model accuracy for test set is 92.4%.

Precision of the model also evaluates through the following equation. precision = TP/TP+FP

$$\text{Precision} = 3662/3662+314 = 0.921$$

5.6.6 Variable Importance in SVM Model

```
> importance = varImp(svm_Linear, scale=FALSE)
> importance
ROC curve variable importance
```

```
Importance
AQ      0.8488
AK      0.8462
AH      0.8247
AR      0.7784
AL      0.7380
AM      0.7177
AD      0.5955
AB      0.5664
AC      0.5430
AO      0.5318
AF      0.5316
AG      0.5289
AN      0.5150
AE      0.5136
AJ      0.5098
AI      0.5022
AP      0.5000
```

Fig. 5.33: variable importance ROC Curve in SVM model snap shot

```
> plot(importance)
```

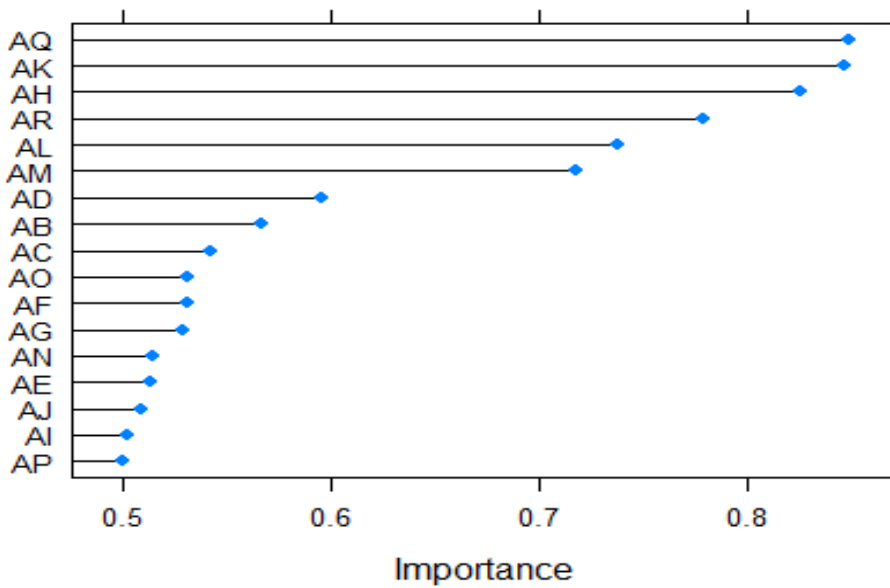


Fig. 5.34: variable importance plot in SVM model snap shot

The above figure 5.32 and 5.33 shows the SVM model variable importance in percentages, according to this, the top four massively importance attributes are AQ (MFIs type), AK (product type), AH (percentage of collateral ownership) and AR (interest rate).

5.7 Summary of Main Findings

The following section presents the summary of the main findings of this work. For easy readability it is presented in table form. As stated earlier (in chapter One) the main objective of building the prediction model is to come up with a pattern for each borrower loan status (Active, Defaulter) that would help in predicting the likely status of a new borrower in terms of these attributes. The experiment for the study is extensively tested on the data set. The dataset in which testing is performed are those which are described in the data preprocessing section. As it is described above, the MFIs data set contain 37380 out of it 20% of the data set is taken as test sample by keeping the remaining 80% of it as training.

The prediction is checked on KNN, SVM, Naïve Bayes and logistic regression algorithms separately, and the result is presented & discussed as follows. The summary of the result also includes confusion matrix, which is a table layout to visualize the performance of a model. A typical confusion matrix consists of rows and columns where each column represents the number of instances in the predicted class and each row represents the number of instances in an actual class. In predictive analytics, a confusion matrix represents the total number of true positives, false positives, false negatives and true negatives.

5.7.1 Results on KNN (Nearest neighbor algorithm)

A separate check conducted in KNN model by changing the value of k to 192,193,194 and 185-205 is done during experimentation, and the percentage of correctly identified loan status among the total is taken as result of the study. Performance of the study is checked through accuracy and precision using confusion matrix. The result obtained from the experiment is stated in the following manner. The following table gives the summary of the four-experiment conducted in KNN model.

no	Experiment label	Attribute used	K value	Accuracy level	Precision
1	Experiment one	18	193	99.91	1
2	Experiment two	18	192	99.87	1
3	Experiment three	18	194	99.87	1
4	Experiment four	18	185-205	Average of (99.89)	

Table 5.1 Summary of KNN experiments result

These accuracy values are very good as stated in [65] [66] estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

5.7.1.1 Confusion Matrix of KNN Classifier for loan status Prediction: -

KNN (Experiment one with value of K 193)		Predicted class	
		Active	defaulter
Actual class	Active	7418	0
	defaulter	7	555

Table 5.2 Confusion matrix result of KNN experiment one

KNN (Experiment two with value of K 192)		Predicted class	
		Active	defaulter
Actual class	Active	7418	0
	defaulter	10	552

Table 5.3 Confusion matrix of KNN experiment two

5.7.1.2 Attribute Contribution for KNN Prediction Model

According to the KNN prediction model, as stated in the above figure 5.17, only eleven attributes are listed as important attributes out of 17 independent attributes, out of these eleven

importance attributes two attributes those are AQ (MFIs organization Type attributes) and AG (sector attributes) contributed massively in the prediction. with the result shows below in descending order in fig 5.35. As a result of this, the results analysis and discussion on this section would be made around the two most contributing attributes.

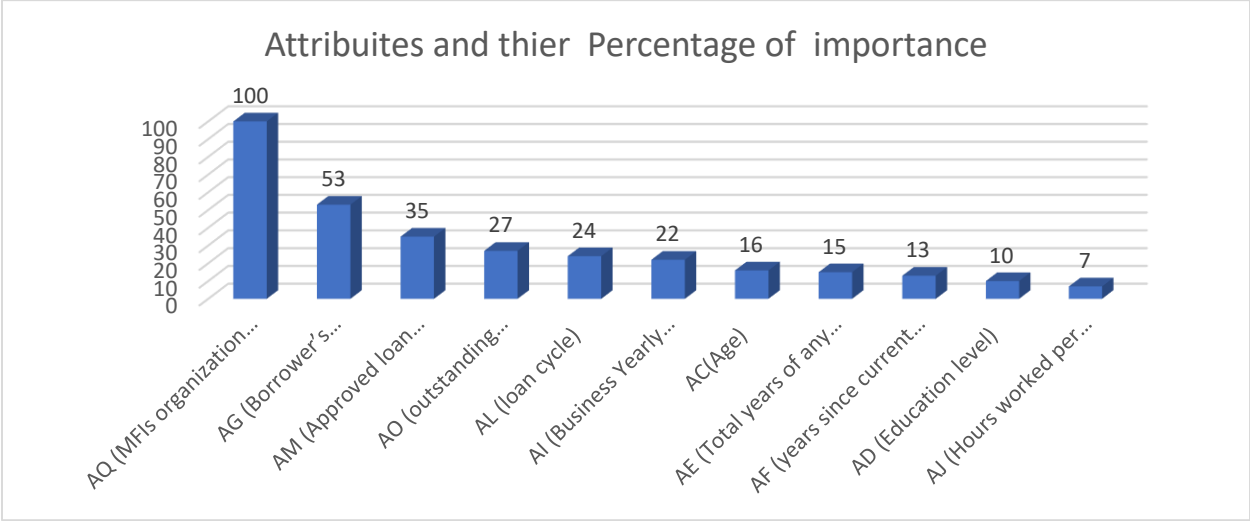


Figure 5. 35: Attributes and their percentage of contribution to the KNN prediction model

5.7.1.2.1 MFIs Type attributes

All the prospects contacted were grouped into three different categories for the MFIs organization type that is NGO affiliated MFIs – (green), governmental affiliated MFIs (orange), fully commercial MFIs (blue) representation. As we see in the prediction accuracy of KNN, it was 99.9% that is very close to 100%, so based on this accuracy we analyses the results, it indicated that most of the borrowers (22127) data contacted had NGO affiliated MFIs, and the rest are having 4179 and 11074 MFIs are government and fully commercial MFIs respectively. The data of MFI type respectively as indicated below in figure 5.36.

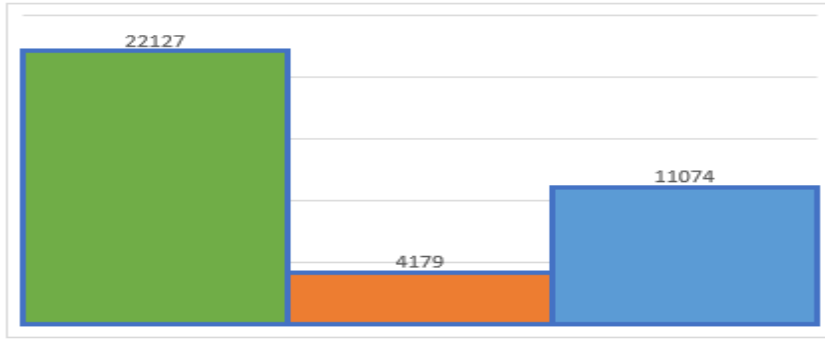


Figure 5.36: Total prospective borrowers in MFIs type

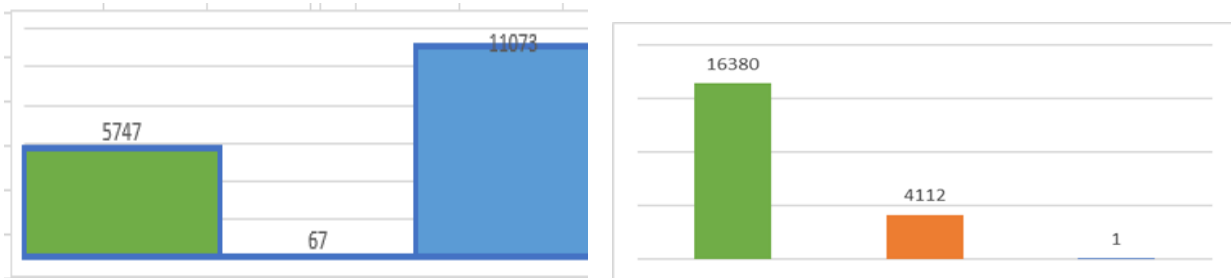


Figure 5.37: Active borrowers in MFIs type

Figure 5.38: defaulter borrowers in MFIs type

Figure 5.37 and 5.38 indicates majority of the loan borrowers who responded the loan effectively (Active borrowers) are fully commercial MFIs, 11073 are active borrowers out of 11074. And governmental type MFIs 4112 borrowers are defaulter out of 4179, only 67 borrowers out of 4179 are active. The case of NGO affiliated MFIs also only 5747 are active out of 22127 borrowers. This means that most of loan borrowers of government and NGO affiliated Ethiopian MFIs are defaulter, and that the MFIs can focus if the borrowers are former customers of both NGO and governmental affiliated MFIs for loan marketing.

5.7.1.2.2 Borrower's Business Sector Attributes

Under the borrower's business sector attribute there were four categories. These were manufacturing (blue), service (red), trade (green) and others (deep green). As depicted by figure 5.39 identified the business sector group with the highest contacts as trade (green) with total number of 16100 contacts followed by service with total number of 11760 contacts 8540 and 980 contacts.

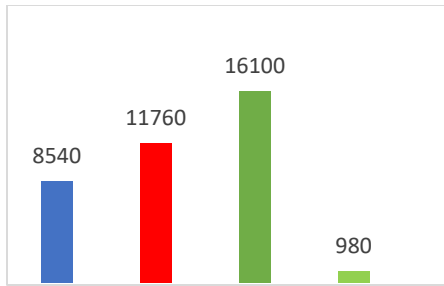


Figure 5.39: data on business sector

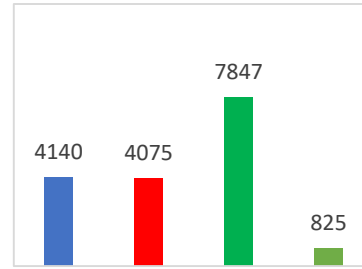


Figure 5.40: defaulter borrowers based on sector

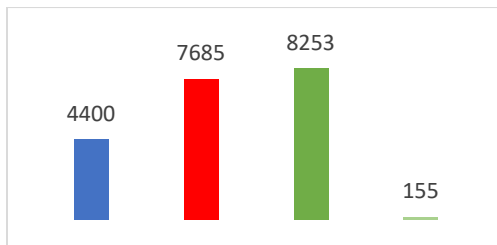


Figure 5.41: active borrowers based on sector

In figure 5.40 and 5.41, it was realized that the service group that responded massively to the loan offer was service group with 7685 Active and 4075 Defaulter responses. The second group that also with 8253 Active and 7847 defaulter response is trade. Again, the third group manufacturing that also responded 4140 Active and 4400 Defaulter responses. Therefore, it can be analyzed that most of the prospective borrowers who responded to the offer are in the service sector attributes, so that the Ethiopian MFIs could target this service business sector for loan marketing, and also can make it very attractive and advantages investing loan business in service sector to the local borrower's people.

5.7.2 Results on SVM (Support Vector Machine)

After the development of the SVM classification model, a number of evaluation techniques were implemented. Confusion matrix is used for evaluate model accuracy and crossable used to visualize the result, and also precision of the model calculated using equation. The SVM classification model was developed to identify prospective Active and Defaulter borrowers in Ethiopia MFIs, it yielded with accuracy values of 92.4%. These accuracy values are very good as [65] [66] estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

5.7.2.1 Confusion Matrix of SVM Classifier for loan status Prediction

SVM		Predicted class	
		Active	defaulter
Actual class	Active	3662	254
	defaulter	314	3246

Table 5.4 Confusion matrix of SVM experiment

5.7.2.2 Attribute contribution for SVM prediction model

According to the SVM prediction model, four attributes that contributed massively (>75%) in the prediction are AQ (MFIs organization Type attributes), AK (Loan product), AH (Percentage of collateral ownership) and AR (interest rate). The result of attribute importance presents in fig 5.33 and fig 5.34 above hear showed below in descending order in fig 5.42. As a result of this, the results analysis and discussion on this section would be made around the four most contributing attributes.

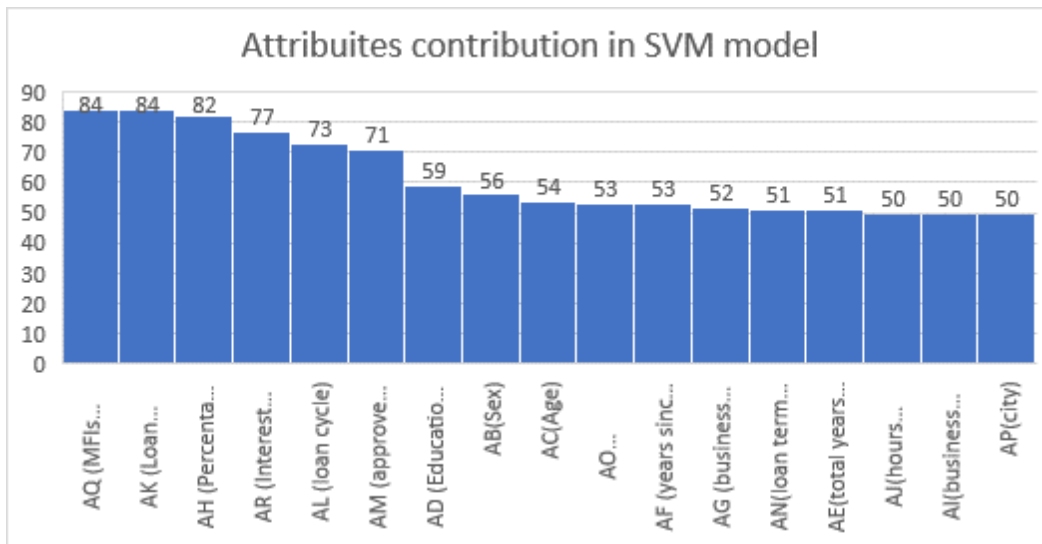


Figure 5.42: Attributes and their percentage of contribution to the SVM prediction model

5.7.2.2.1 MFIs Type Attributes

The loan borrower's in NGO affiliated MFIs (5041 out of 7980 test data set) predicted (43%) defaulter and the rest (57%) as active with 92% accuracy. Similarly, the loan borrower's in government affiliated MFIs (591 out of 7980 test data set) predicted only (8%) Active. And loan borrower's in fully commercial MFIs (2348 out of 7980 test data set) predicted with accuracy of 92.4%, only 4% are defaulter. So, it shows that most of borrower's loan history from fully commercial MFIs are trustworthy than NGO and government affiliated MFIs.

5.7.2.2.2 Loan Product Attributes

The prospective loan product attributes that patronized the loan status mostly is Agricultural product (2980 out of 7980 test data set) with (92%) Active and (8%) Defaulter responses with 92.4% of accuracy. The next interesting finding is that, WEDEP loan product (2939 out of 7980 test data set) had (74%) Active and (26%) Defaulter responses. Micro loan product recorded (1142 out of 7980 test data set) (87%) Defaulter and only (13%) Active responses. General loan product recorded (61 out of 7980 test data set) 98% of it is classified as Active with 92.4% accuracy. And Small loan product type (858 out of 7980 test data set) (61%) defaulter and only (39%) Active responses. So Ethiopian MFIs could target Agricultural, WEDEP and General product borrowers is advantages and little risk than micro and small loan product borrower's for granted loan.

5.7.2.2.3 Percentage of Collateral Ownership Attributes

The loan applicants who had their percentage of collateral ownership are 100% that is 2736 out of 7980 test data classified ,68% as defaulter and the rest 32% of it as Active with 92.4% accuracy. Again, borrower's their percentage of collateral ownership are 50% that is 5244 out of 7980 test data are classified, 18% as defaulter and the rest 82% as Active. As a result of this, it can be analyzed that most of the prospective borrowers with 50% collateral ownership are more trustworthy borrower's in terms of returning back the loan than prospective borrowers with 100% collateral ownership borrowers. Ethiopian MFIs could target this kind of borrowers (50% collateral ownership) business group for loan marketing to invest loan business.

5.7.2.2.4 Interest Rate

Flat interest rate loans are calculated interest based on the amount of money a borrower receives at the beginning of a loan and declining interest rate interest is calculated in MFIs every month on the outstanding loan balance as reduced by the principal repayment every month. The MFIs flat interest type attributes (5620 test data out of 7980 records) classified 57% as defaulter and the rest 43% classified as active with 92.4% of accuracy. Again, the MFIs declining interest type attributes ((2360 test data out of 7980 records)) classified 82.7% classified as active and 17.3% classified as defaulter. As a result of this it shows that MFIs flat interest type is one of the main causes for loan borrowers to be a defaulter.

5.7.3 Results on Logistic Regression Model

Like the experiment held on using nearest neighbor algorithm and SVM, conducting the experiment through MFIs data set with 37380 observations and 18 attributes. The model performance evaluates through accuracy and precision, sensitivity and specificity. Confusion matrix is used for evaluate model accuracy and crossable used to visualize the result. And also, precision of the model calculated using equation. The logistic regression model was developed to identify prospective Active and Defaulter borrowers in MFIs like in other models. In this study also the imbalance data and balanced data is tested. It yielded with accuracy values of 92.4% and 93.8%. These accuracy values are very good as [65] [66] estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

5.7.3.1 Confusion Matrix of Logistic Regression Classifier for loan Status Prediction

Logistic Regression (with balanced data)		Predicted class	
		Active	defaulter
Actual class	Active	3583	173
	defaulter	393	3327

Table 5.5 Confusion matrix of logistic regression experiment

Logistic Regression (with imbalance data)		Predicted class	
		Active	defaulter
Actual class	Active	3692	173
	defaulter	284	3327

Table 5.6 Confusion matrix of logistic regression experiment

5.7.3.2 Attribute contribution for logistic regression prediction model

As shown in figure 5.22, in logistic regression model eight attributes for the statistically significant variables are, AL (loan cycle attributes), AM (approved loan amount attributes), AN (loan term in month attributes), AO (Outstanding principal balance).AC(age), AG (business sector attributes), AI (Business Yearly Earnings attributes) and AD (Education status attributes). These eight attributes have the lowest p-value that indicate a strong association of those attributes to the loan with the probability of having active borrowers.

5.7.4 Results on Naive Bayes Model

For the Naive Bayes prediction model, we use samples of about 37380 loan borrowers' details with 17 independent and one dependent attributes. After the development of the model, evaluation techniques were implemented. Confusion matrix is used for evaluate model accuracy and crossable used to visualize the result, and also precision of the model calculated using equation. The model accuracy is 83.9%, these accuracy values are very good as [65,66] estimates high accuracy to be in the range of 0.75 and 1 representing 75% and 100% respectively.

5.7.4.1 Confusion Matrix of Naive Bayes Classifier for loan status Prediction

Naive Bayes		Predicted class	
		Active	defaulter
Actual class	Active	2772	0

	defaulter	1204	3500
--	-----------	------	------

Table 5.7 confusion matrix of Naive Bayes experiment

5.7.5 Machine Learning Models Comparison

The below (table 5.8) shows a summary of the result obtained from all the four models when all models are trained on around 29400 instances. It has been found that KNN resulted into highest accuracy. But at the prediction level Logistic regression, SVM and Naive Bayes model performed well in terms of accuracy and precision. These machine learning models evaluations(comparison) are done through the model accuracy and precision. The four Machine learning models (SVM, logistic regression, Naive Bayes, and KNN) were used to predict the Ethiopian MFI data on the 18 selected attributes.

Table 5.8 model comparison

	SVM	KNN		Logistic regression		Naive Bayes
		KNN (Experiment one)	KNN (Experiment two)	with balanced data	With imbalance data set	
Total number of instances (test data set)	7980	7980	7980	7476	7476	7476
Correctly classified instance	7973	7973	7970	6910	566	6272
Incorrectly classified instance	7	7	10	7019	557	173
Accuracy	92.4	99.91	99.87	92.89	93.8	83.89
Precision	0.921	1	1	0.9	0.928	1

As stated in above table machine learning algorithm K-NN recorded an extremely higher accuracy (99.9%) than the rest three models but Logistic Regression (93.8%) and SVM (92.4%) show a comparable performance. The result of the study is presents to the domain experts those are senior MFIs managers and experts and they proved technical evaluation to assurance of the results of the study and quality of the data.

5.8. Summary

In this chapter, the processes involved in building four prediction model using SVM, KNN, Naïve Bayes and logistic regression machine learning algorithms as well as the performance evaluation procedures were discussed. The cross table was presented to visualize the model result and the confusion matrix was presented for the accuracy, also precision was calculated.

The discussion of the result in this chapter also about the four-prediction model that were generated. In relation to research question two, the KNN, SVM, Logistic regression and Naïve Bayes classifier was modeled to find out how machine learning can be utilized in making decision whether to extend loan or not, through classify and predict the credit worthiness of loan applicants. Again, in relation to research question 2.2, Which model is more accurate for predicting loan risk through classify defaulter and active loan applicant, the best performance was achieved by KNN, while Naïve Bayes model had relatively the lower performance on the Ethiopia MFIs data set. Logistic Regression and SVM show a comparable performance. In addition, relation to research question 2.1, that is attributes most useful in predicting MFIs loan risk are AL(loan cycle attributes), AM(approved loan amount attributes), AN(loan term in month attributes), AO(Outstanding principal balance).AC(age), AG(business sector attributes), AI(Business Yearly Earnings attributes) and AD(Education status attributes) in logistic regression model and according to SVM model, AK (Loan product), AH (Percentage of collateral ownership),AQ (MFIs organization Type attributes) and AR(interest rate) and also AQ (MFIs organization Type attributes) and AG (sector attributes) in KNN model. The next chapter is about the conclusion, recommendation and future works.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

This chapter summarizes the entire findings of the study. It is divided into two main sections. Again, it provides the conclusion and recommendations and the future works.

6.1 Conclusion

The main objective of this research was to determine how machine learning model could be used to detect the credit worthiness or assess the loan risk of Ethiopian MFIs loan applicant. The necessary data for the experiment was obtained from the selected seven microfinance institutions in Ethiopia, those are Aggar MFI, Vision fund MFI, Oromia MFI, Wosasa MFI, Nisir MFI, Harbu MFI and Pease MFI on a scattered excel sheets which totally amounted to a dataset of 37723. The necessary preprocessing activities were applied on the dataset after which 37380 data was prepared for the experimentation. Classification and prediction model building was experimented with KNN, SVM, Naïve Bayes and logistic regression algorithm. And the tools were used to simulate all the experiment is RStudio using R programming. The confusion matrix was used and then accuracy, and sensitivity were calculated. The three models (KNN, SVM and logistic regression) yielded remarkable results when it comes to correctly classifying instances, KNN with accuracy of 99.91%,99.87%,99.874% in the first, second and third experiments respectively, SVM with accuracy of 92.4%, and logistic regression with accuracy of 92.8 and 93.8 on balanced and imbalanced data set respectively. According to this experimentations the attributes AL(loan cycle attributes), AM(approved loan amount attributes), AN(loan term in month attributes), AO(Outstanding principal balance).AC(age), AG(business sector attributes), AI(Business Yearly Earnings attributes) and AD(Education status attributes) in logistic regression model and according to SVM model, AK (Loan product), AH (Percentage of collateral ownership),AQ (MFIs organization Type attributes) and AR(interest rate) and also AQ (MFIs organization Type attributes) and AG (sector attributes) in KNN model, are found to be relevant (important) predictors for the target class borrower's loan status (active and defaulter).

From the results of the experiments it can be concluded that the Machine learning algorithms and techniques can be effectively applied on the microfinance in order to generate loan risk predictive models with an acceptable level of accuracy.

Therefore, MFIs CEO, MFIs operation department managers, loan officers, MFIs Information technology services divisions in Ethiopian microfinance institutions and also National bank of Ethiopia microfinance supervision division and other financial institutions in Ethiopia can apply these Machine learning models to detect possible defaulter and active loan applicants before granting loan to the borrowers. Also, they can use the models to automate the process of granting loans in MFIs, so as to limit or to completely end from granting loans to risky loan borrowers. By adopting this way, it is expected that MFIs in Ethiopia would operate very well to granting loan for their borrower's and will be able to make the expected returns.

However, since the quality and size of dataset used, in addition to the Machine learning tools and techniques are essential factors for the modeling performance, an increased size in the dataset (with increased and different attributes and include all the MFIs loan product type in Ethiopia) could result in an improved modeling. It could have enabled the research to make use of larger data with more attributes than those used in this study and also address other main risk in Ethiopian microfinance institutions like operational risk and market risk. Because of time and other constraints, the model building experimentation conducted is based on selected seven microfinance institutions with just a 37380 borrower's data but it is the researcher's belief that it would have resulted in improved model if it includes all 35 MFIs in Ethiopia with big data size and all type of loan products borrower's, and also through other techniques were also utilized. Hence following section presents some recommendations made based on the result of the research.

6.2 Recommendations and Future Works

Even though the investigation undertaken is mainly for academic purpose, it will have important contribution for the microfinance institutions and for other researchers interested in similar area. Although the results of this study are inspiring, there are problem areas that need further investigation for future work to attain better inclusive model and also bring it to an

operational level. Therefore, the researcher forwards the following issues as a future research direction based on this study: -

- As discussed in the paper we only explore and demonstrated that Machine learning algorithms can be used to limit or possibly completely extinguish to the prediction of loan risk in Ethiopian MFIs. But there are generally four main risks, loan risk, operational risk, portfolio risk and marketing risk, that exist in every microfinance and other financial institutions. Therefore, interested researchers can look into other risk in microfinance that includes portfolios, operation and marketing risk from machine learning perspective.
- Different classification algorithms such as neural network, decision tree and ensemble learning can be employed for Ethiopian MFI loan risk assessment with increased and different attributes that include all the MFIs with their all loan products type borrower's by different researchers to see if it could result in a different.

REFERENCES

- [1] Kiflie Hayleeyesus, "The Impact of Microfinance Institutions on Poverty Alleviation (A Case Study in Ethiopia)" MBA Thesis, Ritsumeikan Asia Pacific University, September 2016.
- [2] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo," Consumer Credit Risk Models via Machine-Learning Algorithms," MIT Elsevier B.V,May 9, 2010.
- [3] Martina Sandberg," Credit Risk Evaluation using Machine Learning," Linköping University, September 2017.
- [4] Lkhagvadorj Munkhdalai, Oyun-Erdene Namsrai, Keun Ho Ryu1," Credit Scoring with Deep Learning", International Conference on Information, System and Convergence Applications. February 2018.
- [5] Bolarinwa Akindaini, Martti Juhola," Machine Learning Applications in Mortgage Defaults Predictions" University of Tampere, November 2017.
- [6] Jozef Zurada, Martin Zurada," How Secure Are "Good Loans": Validating Loan-Granting Decisions and Predicting Default Rates on Consumer Loans", the Review of Business Information Systems, Volume 6, May 2011.
- [7] Amar Nath Jha, Soumitro Chakravarty," Viability of "Credit Scoring in Microfinance" for Developing Countries", International Review of Social Sciences and Humanities Vol. 3, No. 1 (2012), pp. 104-107.2012.
- [8] Acheampong Amponsah," Enhancing Direct Marketing and Loan Application Assessment Using Data Mining," Kwame Nkruma University, Kumasi, Ghana, April, 2016.
- [9] Hongri Jia," Bank Loan Default Prediction with Machine Learning" passion for data science, Apr 10, 2018.
- [10] Ebisa Deribie, Getachew Nigussie and Fikadu Mitiku," Filling the breach: Microfinance" Journal of Business and Economic Management 1(1): 010-017, January 2013.
- [11] C. Serrano-Cinca, B. Gutiérrez-Nieto and N. M. Reyes," A Social Approach to Microfinance Credit Scoring" Université Libre de Bruxelles, Research Institute in Management Science, 2013.

- [12] Sydney Chikalipah, Credit risk in microfinance industry: Evidence from sub-Saharan Africa, *Review of Development Finance* 8 (2018) 38–48, 2 June 2018.
- [13] Hans Delliën, Mark Schreiner, “Credit Scoring In Microfinance” Vol. 1, No. 2, October 2003 Women’s World Banking.
- [14] Marguerite Berger, “Microfinance: An Emerging Market within the Emerging Markets”, Inter-American Development Bank Social Sustainable Department Micro, Small and Medium Enterprise Division, Washington, DC 2008.
- [15] Ansen Mathew, “Credit Scoring Using Logistic Regression” Master's Theses, San Jose State University Spring May, 2017.
- [16] Ram Babu, Rama Satish, “Improved of K-Nearest Neighbor Techniques in Credit Scoring,” *International Journal for Development of Computer Science & Technology*, Volume-1, Issue-2, Feb-March-2013.
- [17] Abbas Keramati, Amin Omidvar, “Default Probability Prediction of Credit Applicants Using a New Fuzzy KNN Method with Optimal Weights,” IGI Global, ch024, 2015.
- [18] Jesper De Groot, “Credit risk modeling using a weighted support vector machine” Utrecht University, Master Thesis, September 23, 2016.
- [19] Olatunji J. Okesola, Kennedy O. Okokpujie, Adeyinka A. Adewale, Samuel N. John, Osemwegie Omoruyi, “An improved Bank Credit Scoring Model a Naïve Bayesian Approach,” *International Conference on Computational Science and Computational Intelligence*, 2017.
- [20] Aida Krichene, “Using a naive Bayesian classifier methodology for loan risk assessment,” *Journal of Economics, Finance and Administrative Science* Vol.22No.42, 2017 pp. 3-24.
- [21] Tony Van Gestel, Bart Baesens, Dr. Ir. Joao Garcia, Peter Van Dijke, “A Support Vector Machine Approach to Credit Scoring,” *impact research, Credit Methodology Global Market Risk*, Dexia Group, 2015.

- [22] D. L. Gupta, A. K. Malviya, Satyendra Singh,” Performance Analysis of Classification Tree Learning Algorithms,” International Journal of Computer Applications (0975 – 8887) Volume 55– No.6, October 2012.
- [23] Mark Stamp,” A Survey of Machine Learning Algorithms and Their Application in Information Security: An Artificial Intelligence Approach,” San Jose State University, San Jose, California, September 2018.
- [24] Kajaree Das, Rabi Narayan Behera” A Survey on Machine Learning: Concept, Algorithms and Applications,” International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017.
- [25] Alex Addae-Korankye,” Causes and Control of Loan Default/Delinquency in Microfinance Institutions in Ghana” American International Journal of Contemporary Research Vol. 4, No. 12; December 2014.
- [26] C. Serrano-Cinca, B. Gutiérrez-Nieto,N. M. Reyes,” A Social Approach to Microfinance Credit Scoring” Brussels School of Economics and Management Centre Emile Bernheim, CEB Working Paper No. 13/013 2013.27
- [27] Marguerite Berger,” Microfinance: An Emerging Market within the Emerging Markets” World Scientific and Imperial College Press, Washington, DC 2000.
- [28] Edward Yeallakuor Baagyere,Regina Esi Turkson,”A Machine Learning Approach for Predicting Bank Credit Worthiness” University of Electronic Science and Technology of China ,2016 IEEE
- [29] M.A.R. Khalid,M.A.H. Farquad,” Comparative Analysis of Support Vector Machine: Employing Various Optimization Algorithms” 14th International Conference on Information Technology,2015 IEEE
- [30] Jorma Laaksonen,Erkki Oja,” Classification with Learning k-Nearest Neighbors”, Helsinki University of Technology Laboratory of Computer and Information Science, ©1996 IEEE

- [31] Anjali Ganesh Jivani," The Novel k Nearest Neighbor Algorithm" 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 – 06, 2013, 2013 IEEE.
- [32] Rahul Saxena," Introduction to k Nearest Neighbor Classification and Condensed Nearest Neighbour Data Reduction" Data Science, Machine Learning journal monthly blog , December 23, 2016.
- [33] Yuguang Huang,Lei Li," Naïve Bayes Classification Algorithm Based on Small Sample Set" Beijing University of Posts and Telecommunications, Proceedings of IEEE CCIS2011.
- [34] Jiangtao Ren, Sau Dan Lee, Xianlu Chen," Naive Bayes Classification of Uncertain Data" 2009 Ninth IEEE International Conference on Data Mining.
- [35] Ashlesha Vaidya," Predictive and Probabilistic Approach Using Logistic Regression:Application To Prediction of Loan Approval" 8th ICCNT 2017 July 3-5, 2017, IIT Delhi. IEEE – 40222.
- [36] sheena Angra,sachin Ahuja, "Machine Learning and its applications: A Review" Curin chitkara university ,2017 IEEE
- [37] Simret Solomon,Tibebe Beshah," Predicting Customer Loyalty Using Data Mining Techniques" University of South Africa (UNISA), Addis Ababa, Ethiopia,2014
- [38] Aida Krichene Abdelmoula," Bank credit risk analysis with k-nearestneighbor classifier: Case of Tunisian banks" Accounting and Management Information Systems Vol. 14, No. 1, pp. 79-106, 2015.
- [39] Vimala S., Sharmili K.C," Prediction of Loan Risk using Naive Bayes and Support Vector Machine" International Conference on Advancements in Computing Technologies - ICACT 2018.
- [40] Paulius Danenasa, Gintautas Garsvab, Saulius Gudasc," Credit Risk Evaluation Model Development Using Support Vector Based Classifiers" International Conference on Computational Science, ICCS 2011.
- [41] Thomas G. Dietterich," Ensemble Learning" Cambridge, MA: The MIT press,2002.
- [42] Anchal Goyal, Ranpreet Kaur," Loan Prediction Using Ensemble Technique" International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

- [43] Yebabie Yoseph,” Assessment of Credit Risk Management in Micro Finance Institutions” Master thesis Business Administration in Accounting and Finance, St. Mary’s University Ethiopia,2017.
- [44] Hable Asrat” Assessment of Credit Management Practice at United Bank S.C” Master thesis Business Administration in finance, Addis Ababa University,2018.
- [45] Zhi-Hua Zhou,” Ensemble Learning” National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China,2016.
- [46] Sara Worku, “Application of Data Mining Technology for Credit Risk Assessment in Addis Credit and Saving Institution, “Master’s Thesis, Addis Ababa University, June 2016.
- [47] JI JIAQI,Yeongjee Chung,” Research on K Nearest Neighbor Join for Big Data” International Conference on Information and Automation (ICIA) Macau SAR, China, July 2017, Proceedings of the 2017 IEEE.
- [48] Samuel Setargie,” Credit Default Risk and its Determinants of Microfinance Industry in Ethiopia” Aksum University,College of Business and Economics,2013.
- [49] S. Rajasekar, P. Philominathan, V. Chinnathambi,” Research Methodology” October 2013.
- [50] Tesfaye Sisay,” Factors Influencing Loan Repayment Performance Micro and Small Enterprise Borrowers Financed by Micro and Small Enterprise Borrowers Financed by Microfinance Institution.” MBA Thesis, Addis Ababa University College of Business and Economics, January 2018.
- [51] Haidar Osman, Mohammad Ghafari, Oscar Nierstrasz,” Automatic Feature Selection by Regularization to Improve Bug Prediction Accuracy”, MaLTeSQuE 2017, Klagenfurt, Austria ,2017 IEEE.
- [52] Raheel Shaikh, ”Feature Selection Techniques in Machine Learning ” Towards Data Science, Oct 28, 2018.
- [53] Raschka, Sebastian,” Predictive modeling, supervised machine learning, and pattern classification” 2014 IEEE.

- [54] R. David A. Dickey," Introduction to Predictive Modeling with Examples," SAS Global Forum, 2012.
- [55] Ritesh Jain," Predictive Modeling for Chronic Conditions", Master's Thesis, College of Computer Science and Engineering, Florida Atlantic University, May 2015.
- [56] Behrouz Shamsaei, Cuilan Gao," Comparison of some machine learning and statistical algorithms for classification and prediction of human cancer type" The University of Tennessee, USA, 2016 IEEE
- [57] Thomas Debray," Classification in Imbalanced Datasets" publication at: researchgate.net/publication/265116459, on September 2014.
- [58] Yebabie Yoseph "Assessment of Credit Risk Management in Micro Finance Institutions" MBA Thesis, St. Mary's University Ethiopia, January 2017.
- [59] Parul Pandey "A Comprehensive Guide to Data Visualization" Towards Data Science Feb 4 2019.
- [60] Saradnici "data normalization" Microsoft document, Microsoft azure, docs.microsoft.com 1st Jan, 2019. Available: <https://docs.microsoft.com>. [Accessed: may. 2, 2019].
- [61] Shweta Taneja, Charu Gupta, Kratika Goyal "An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering" 2014 Fourth International Conference on Advanced Computing & Communication Technologies, © 2014 IEEE.
- [62] Dahee Choi, Kyungho Lee an Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation" Hindawi Security and Communication Networks Volume 2018, Article ID 5483472,25 September 2018.
- [63] Raj Rakshit , Anwasha Khasnobish," A Novel Approach to the Identification of Compromised Pulmonary Systems in Smokers by Exploiting Tidal Breathing Patterns", TCS Research and Innovation, Sensors 2018, 18, 1322, 25 April 2018.
- [64] Fadhl M. Alkawaa, Kumardeep Chaudhary "Deep learning accurately predicts estrogen receptor status" Journal of Proteome Research is published by the American Chemical Society, November 8, 2017.

- [65] François Fous, Marco Saerens” Evaluating performance of recommender systems: An experimental comparison” 2008 IEEE.
- [66] Max Kuhn” Package ‘caret’ Classification and Regression Training book 2019-04-27 04:50:03 UTC ,April 27, 2019.
- [67] John Muschelli” ROC and AUC with a Binary Predictor: a Potentially Misleading Metric”, Johns Hopkins Bloomberg School of science, March 13, 2019.
- [68] Juliana Ivanpáková, František Babip” Comparison of Different Machine Learning Methods on Wisconsin Dataset” IEEE 16th World Symposium on Applied Machine Intelligence and Informatics • February 7-10,2018.
- [69] Steffen Huber, Hajo Wiemer, Dorothea Schneider” DMME: Data Mining Methodology for Engineering Applications – A Holistic Extension to the CRISP-DM Model” 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, Gulf of Naples, Italy ,18-20 July 2018.
- [70] Bruno C. da Rocha and Rafael T. de Sousa Junior,” Identifying Bank Frauds Using CRISP-DM and Decision Trees” International journal of computer science & information Technology (IJCSIT) Vol.2, No.5, October 2010, 162 – 169.
- [71] Sharma B.” Guide to Credit Scoring in R”, Interdisciplinary Independent Scholar in risk management, Sept 23 2009, <https://excite.com> .[Accessed: may. 7, 2019].
- [72] Ana Azevedo, Manuel Filipe Santos” KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW” IADIS European Conference Data Mining,2008.

Appendix I

Sample collected data set from microfinance institutions (Before Preprocessing the data).

	A	B	C	D	E	F	G	H	I	J
1	EnglishFull Name	City	sub city	sex	Age	StartingC apital	Educatio n	TotalYearsOfB usinessExperie nce	YearsInBusine ssEstablished	EverBorr owedMo ney
2	hidden	Addis Ababa	AK Wered	F	32	10000	graguate	8	1	FALSE
3	hidden	Addis Ababa	AK Wered	F	50	5000	Secondary	9	9	FALSE
4	hidden	Addis Ababa	AK Wered	M	25	2000	graguate	2	2	FALSE
5	hidden	Addis Ababa	AK Wered	F	38	50000	graguate	12	6	FALSE
6	hidden	Addis Ababa	AK Wered	F	30	5000	Secondary	6	4	FALSE
7	hidden	Addis Ababa	AK Wered	M	40	3000	Primary	11	8	FALSE
8	hidden	Addis Ababa	AK Wered	M	43	5000	Secondary	20	10	FALSE
9	hidden	Addis Ababa	AK Wered	M	36	50000	Secondary	8	8	FALSE
10	hidden	Addis Ababa	AK Wered	F	47	20000	Secondary	12	12	FALSE
11	hidden	Addis Ababa	Bole Wered	M	42	5840	Secondary	6	10	FALSE
12	hidden	Addis Ababa	Bole Wered	F	40	14009	Secondary	2	3	FALSE
13	hidden	Addis Ababa	Bole Wered	F	49	2000	Secondary	24	1	FALSE
14	hidden	Addis Ababa	Bole Wered	F	22	2700	graguate	1	1	FALSE
15	hidden	Addis Ababa	Bole Wered	F	31	7000	graguate	3	3	FALSE
16	hidden	Addis Ababa	Bole Wered	F	54	50000	Secondary	6	2	FALSE
17	hidden	Addis Ababa	Bole Wered	M	33	1500	graguate	4	4	FALSE
18	hidden	Addis Ababa	Bole Wered	M	40	22090	graguate	4	4	FALSE
19	hidden	Addis Ababa	Bole Wered	F	27	5000	graguate	6	1	FALSE
20	hidden	Addis Ababa	Bole Wered	F	37	5002	Secondary	6	10	FALSE

K	L	M	N	O	P	Q	R	S	T	U
Sector	SubSector	Business License	TINNo	PercentageOfOwnership	BusinessYearlyEarnings	HoursWorkedPerWeek	TrainingType	Telephone	IDNumber	loan status
Manufact	Textile an	TRUE	3757073	100	1200000	56	Basic	9.12E+08	1021010144	Active
Manufact	Intermedi	TRUE	3067873	100	500000	84	Basic	9.11E+08	1021030153	Active
Manufact	Contractir	TRUE	39419967	100	450000	48	Basic	9.13E+08	1022010148	Active
Trade	Retail	TRUE	5976382	100	190000	54		9.12E+08	1023020146	Active
Trade	Retail	TRUE	14471425	100	60000	78	Basic	9.13E+08	1023020147	Active
Trade	Retail	TRUE	823686	100	500000	72	Basic	9.21E+08	1023020149	Active
Trade	Retail	TRUE	143916	100	300000	54	Basic	9.12E+08	1023020150	Active
Manufact	Small tran	TRUE	2312006	100	25000	12	Basic	9.11E+08	1024010145	Active
Other	Other	TRUE	3023814	100	2000000	12	Basic	9.12E+08	1026010151	Active
Manufact	Textile an	TRUE	3121023	100	300000	48	Basic	9.11E+08	1071010454	Active
Manufact	Textile an	TRUE	45552663	100	150000	48	Basic	9.61E+08	1071010479	Active
Manufact	Textile an	TRUE	3463111	100	1000000	48	Basic	9.29E+08	1071010504	Active
Manufact	Intermedi	TRUE	45065894	100	150000	48	Basic	9.11E+08	1071030439	Active
Manufact	Intermedi	TRUE	6015873	100	300000	48	Basic	9.11E+08	1071030452	Active
Manufact	Intermedi	TRUE	7058147	100	18000	12	Basic	9.11E+08	1071030507	Active
Manufact	Intermedi	TRUE	26618504	100	10000	48	Basic	9.36E+08	1071030521	Active
Manufact	Wood wor	TRUE	9652465	100	240000	48		9.11E+08	1071050460	Active
Manufact	Intermedi	TRUE	24611505	100	100000	48	Basic	9.13E+08	1071060451	Active
Manufact	Intermedi	TRUE	4124109	100	100000	48	Advanced	9.12E+08	1071060453	Active

V	W	X	Y	Z	AA	AB
loan product	loan cycle	approved	loan term in month	outstanding principal balance	branch name	MFIs Type
WEDEP	2	100000	12	66000	Addis Aba	oromia
WEDEP	4	150000	12	99000	Addis Aba	oromia
WEDEP	2	700000	12	462000	Addis Aba	oromia
WEDEP	2	150000	12	99000	Addis Aba	oromia
WEDEP	2	5000	12	3300	Addis Aba	oromia
WEDEP	1	10000	12	6600	Addis Aba	oromia
WEDEP	2	25000	12	16500	Addis Aba	oromia
WEDEP	2	13000	12	8580	Addis Aba	oromia
WEDEP	1	100000	12	66000	Addis Aba	oromia
WEDEP	3	20000	12	13200	Addis Aba	oromia
WEDEP	2	8000	12	0	Addis Aba	oromia
WEDEP	2	200000	12	0	Addis Aba	oromia
WEDEP	1	20000	12	4997.04	Addis Aba	oromia
WEDEP	3	25000	12	20832	Addis Aba	oromia
WEDEP	3	8000	12	3998	Addis Aba	oromia
WEDEP	4	25000	12	16664	Addis Aba	oromia
WEDEP	1	8000	12	4665	Addis Aba	oromia
WEDEP	3	8000	12	1995	Addis Aba	oromia
WEDEP	1	6000	12	0	Addis Aba	oromia

Appendix II

RStudio R code used to plot attributes.

```
> EMFILOAN = read.csv(file.choose() , sep = ',')
> View(EMFILOAN)
> colors = c("green", "blue", "red", "violet", "orange","yellow", "pink", "cyan")
> plot (EMFILOAN$loan.status,EMFILOAN$loan.term.in.mounth,col=colors,main="loan status based o
n loan term",xlab="loan status",ylab="loan term")
> plot(EMFILOAN$loan.status,EMFILOAN$HoursWorkedPerWeek,col=colors,main="loan status based
on loan cycle",xlab="loan status",ylab="loan cycle")
> plot(EMFILOAN$loan.status,EMFILOAN$sex,col=colors,main="loan status based on loan cycle",xlab
="loan status",ylab="loan Product(WEDEP,Micro,small,Agri,Gen)")
> plot (EMFILOAN$loan.product,EMFILOAN$loan.status,col=colors,main="loan status based on loan p
roduct",xlab="loan status",ylab="loan Product(WEDEP,Micro,small,Agri,Gen)")
> plot (EMFILOAN$loan.status,EMFILOAN$Education,col=colors,main="loan status based on Educatio
n",xlab="loan status",ylab="education(Graguate,pri,High,No Edu)")
```

