Predicting Probability of Credit Card Default at the Stage of Credit Card Application Using Supervised Machine Learning Approaches

A dissertation submitted for the Degree of Master of Business Analytics

S.D.P.A Amarasinghe

University of Colombo School of Computing

2019

Predicting Probability of Credit Card Default at the Stage of Credit Card Application Using Supervised Machine Learning Approaches

By

Pasan Anuradha Amarasinghe

Reg No: 2018/BA/003

Index No: 18880031

A dissertation submitted to the
University of Colombo School of Computing
In partial fulfillment of the requirements
Of the Master of Business Analytics Degree

University of Colombo School of Computing
University of Colombo, Sri Lanka
2019

DECLARATION

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: S.D.P.A Amarasinghe

Registration Number: 2018/BA/003

Index Number: 18880031

Pasan Amarasinghe

Signature   Date: 09.09.2021

This is to certify that this thesis is based on the work of

Mr. S.D.P.A Amarasinghe

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by,

Supervisor Name: Dr. D.A.S Atukorale

Signature   Date: 12/09/2021

Dedication

I would like to dedicate this thesis to my family, my thesis supervisor and all lectures that guided me to this work. A special feeling of gratitude to my loving parents and wife whose words of encouragement and push for tenacity ring in my ears. I also dedicate this thesis to my friends who have supported me throughout the process. I will always appreciate all they have done.

Acknowledgement

Firstly, I would like to thank the all of the persons that supported for making this thesis possible and for seeing me through the course of my studies at the University of Colombo. This thesis would not have been possible without the supervision and assistance of Dr. Ajantha Athukorale who was always patient with me. My heartfelt gratitude goes to all the respondents. To my friends I thank you for the encouragement, motivation and support in all aspects of my life. Finally, to my beloved loving family My parents, My Wife and Kids, I thank you for all the support and life mentoring that has made me the person I am today.

Abstract

Increasing Non-Performing Loans and advances of Banking and Financial institutions have become one of the main problems in this era. Specially, increasing the number of willful defaulters of the banking and finance sector. Credit Card portfolio is representing the major part of loans and advances portfolio of a Bank. Banks are issuing credit cards by analyzing the credit card application. Except for fully secured credit cards, other credit cards are not required to attach any security for recovering the non-performing amount.

Therefore, the Managers and the officers who are issuing credit cards must have better understanding and experience to analyze the credit card applications received by customers. But, most of analysis techniques using by them have more personal judgements rather than analyzing it in a proper manner. Therefore, proper credit card application analysis and customer credit scoring model is need for this type of transactions.

This research study is focusing to develop a machine-learning process using supervised learning methods to predict Credit Card default probability of Credit Card applicants before issuing the Credit Card to customers. This model can apply to the banking sector of Sri Lanka to reduce ability of transferring Credit Cards to Non-Performing section.
The review of the literature guided to identify the previous studied in relating to previously developed models to predict of transferring Non-Performing Credit Cards research problem.

To develop a supervised learning model and testing the model by using the data in relating to Bank of Ceylon credit card client's is used for this study.
The research study will be conducted using supervised learning algorithms Naïve Bayes, decision trees, linear regression, logistic regression, k-nearest neighbor algorithm and support vector machines. After analyzing, using all of these algorithms finally the best algorithm for this kind of prediction will be selected.

Table of Contents

List of Figures

List of Tables

Chapter One

Introduction

1.1 Chapter Introduction

This chapter provides the framework of the overall research study, which is to predict the probability of Credit Card default at the Credit Card application stage using supervised learning approaches. It contained a brief description of the background and scope of the research followed by the study of objective and literature review of the study. This research proposal explicates the context of the client/ customer that benefited of the study, methodology and methods will use to conduct this study accordingly.

1.2 Background of the Study

The global banking and financial services industry have faced a remarkable shift in meeting new challenges. New technological innovations, advanced communication modes, the internet and the expansion in the bank- branch networks lead intense competition in the banking industry. As a result of Sub-Prime Crisis and Bank runs, strict banking rules and regulations have imposed on Banks by national and international Banking authorities. These rules and regulations have further restricted the business opportunities of the Banks. Due to the consequences of Sub- Prime Crisis and the Basel III Capital accord implementation, banks were compelled to have more precautionary actions than ever before in addressing high-risk exposures. Banks have no competitive advantages for a prevailing long time- period, while banks are becoming more demanding and selective in their preferences.

The health of the economy is closely related to the soundness of its financial system. One of the most important participants in the financial system is the banking system. Financial sector of Sri Lanka is dominated by the banking enterprises, especially the commercial banks. For the development of the economy, the banks provide a higher portion of strength. With the developing and modern financial environment, the Banks also improve the quality of the service that they are providing to the customers through new technologies. Sri Lanka is no exception to these effects and almost all industries including private and nationalized banks are providing varied services to attract the customers' community it is treated as assets of the Banks. In Sri Lanka, there are 26 Licensed commercial banks, 7 Licensed specialized banks and 46 Licensed financial companies. Government owned commercial banks hold the dominance of this sector until 80's. Then Private banks gradually entered in to financial system. Even though the nationalized commercial banks are holding most of the market share of the banking industry the recent growth of the private banking performance and efficiency is significant.

The banking industry of Sri Lanka has also undergone tremendous shift due to intense competition and technological developments. Non- Banking Financing Companies and Multinational Banks caused major changes in the ways of running banking businesses in Sri Lanka in the last decade. The banking culture, business strategies and the ways of treating the customers have been changed. The domestic banks have compelled to adjust to the new culture and the new technologies, which have adopted by the foreign banks. Increasing non-performing loans, government intervention of deciding lending rates and the low deposit mobilizations are the features of the banking industry that are that are creating many problems in the industry.

In addition to above problems, the new challenges faced by banks as follows.

- ICT revolution
- Challenges posed by globalization
- Problems with outsourcing
- Race for adoption of new technology

No matter whatever the problems, bankers need to satisfy two opposing objectives of profitability and liquidity.

In Bank's balance sheet deposits represents the Liabilities and the Loans & Advances represents the Assets. Banks always look for the safeguards of the deposits of the customers when increasing the profitability. In order to accomplish the same, Banks must provide the high-quality Loans & Advances that reduce the Non-Performing Advances ratio. When considering Loans & Advances portfolio, it consists with various kind of Loans & Advances specially Term Loans, Credit Cards etc.

Considering the Loans & Advances portfolio, Banks always looking at the Ratio of Non-Performing Loans & Advances (NPL) because it will affect to the profitability of the Bank and safety of the deposits of depositors of the Banks. Credit cards holds a significant portion of credit portfolio of any Bank and financial institutions. Credit cards issued without any security is resulting to increasing the risk of credit card defaults than other loans and advances. Banks offers credit cards considering the credit worthiness and financial stability of the customers of the Bank. Specially referring to the repayment capacity of monthly credit card statement balance and reliable income source of the customers.

Most of researchers argued that mon-performing loans and advances has major impact to the Banks stability as well as it will lead to major economic problems to the overall banking industry of the country. Each of non-performing loans and advance shows the unprofitability of the Bank and financial institution. Reducing the ratio of non-performing loans and advances increase the growth and stability of the financial sector.

In past decade Credit cards become the major consumer lending product following the personal loans in Sri Lankan banking sector. Also, credit card is a most flexible and easiest way to use banks money in short term basis. Using credit cards people pay utility bills, purchase goods and services, implement installment plans and do international transactions easily. But if the customer does not pay the monthly installments of the credit cards, the credit card will be defaulted and it will impact negatively to the Banks. These defaulted credit cards will write offs at some extend and it will reduce the profit of the Banks drastically. These write-offs will result to significant financial losses to the Bank on top and damage the overall credit rating of the Bank as well as customers. Therefore, these non-performing credit cards must handle carefully by the Banks.

If the Banks can predict the most probable default customers accurately, then the Banks can reduce the non-performing credit cards easily. Each and every Bank use various kind of credit default predictions and credit default preventing guidelines to deidentify non- performing credit cards accurately. Even though plenty of solutions to the credit card default predictions using credit card data in many research papers there are few studies that doing for default prediction in application process of the credit cards. This project focus on the identify and predict non-performing credit cards in early stage (application stage) using supervised machine learning approaches.

1.3 Motivation of the Study

In past few years Sri Lankan Banking Sector faced huge losses as a result of non-repayment of loans and advances. Specially, due to the borrowers who are unwilling to repay their loans and advances will result to huge non-performing loans and advances. Specially over 5% NPL ratios that can be seen in many Banks. Credit card defaults has contributed a significant portion of this high non-performing loans and advances.

Major reason for this non-performing credit cards and other loans & advances is inappropriate decision making and granting the credit facilities and issuing credit cards to the customers using un qualified applications. Managers and employees of the Banks must have experience to identify the default customers which has become a major challenge in recent past. Increasing the accumulated amount of usage of credit cards reduce the repayment capacity of customers drastically and it will lead to huge financial losses to the Banks. Due to revolving credit facility, the monthly repayment amount of credit cards will differ month by month. Hence, effective identification and monitoring of these aspects specially in the event of accepting the application of the credit cards reduce the probability of defaults and reduce the losses of the Bank. Incorporating machine learning techniques for the prediction of defaulters can be useful for the Banks in designing preventive measures and thereby avoiding losses.

1.4 Objectives of the Study

To find out the best classification algorithm for predicting Credit Card default in initial stage of the Credit Card process and make model for identify the best customers to grant a Credit Card in Sri Lankan banking sector.

1.5 Research Question

There are many classification learning algorithms that used to predict and select the best customer for approving the Credit Cards and this will result to decrease the non performing Credit Card percentage of the banking system. Further, it is important to study whether the different algorithms behave differently. Therefore, following research question need to be addressed.

"To examine different classification algorithms and identify the best classification algorithm to predict Credit Cards default in the initial stage of the Credit Card processing process."

1.6 Scope of the Study

According to the data in relating to non-performing loans and advances in relating to loans and advances of the banking sector of Sri Lanka drastically increasing during past decade. Non-Performing loans and advances ratios of banks increasing over 5% and it reduce the profitability of the banks and increase the risks towards the banking sector.

Specially in defaults in credit cards are increasing drastically in recent years. Customers are become defaulters of credit cards willingly. This trend affects negatively to the banking sector. Managers and providers of credits cards in banking and finance sector must have ability to identify the credit card defaulters easily. But various Banks use various kind of credit scoring models and risk analyzing models using both statistical and machine learning approaches. In most Sri Lankan banks use manual methods to identify the credit risks of consumers.

Even though there are various kind of statistical and machine learning approaches used to identify the credit risk and defaulters in Sri Lankan banks there are less studies done to identify the credit card defaulters in credit card application process.

This project intends to the fulfill this gap by predicting probability of Credit Card Default at the stage of Credit Card Application using supervised Machine Learning Approaches for Bank of Ceylon.

1.7 Feasibility of the Study

Prediction of Credit Card defaults is not an easy task. Currently in Bank of Ceylon and other commercial banks in Sri Lanka use various kind of methods to identify the Credit Card defaulters. Specially using manual credit scoring models and risk measuring models they identify the possible credit card defaulters.

Machine learning models are used rarely to identify credit card defaulters. Bank of Ceylon has issued over 100,000 credit cards to their customers and maintain database of the credit cards. This database includes both performing and non-performing credit cards. Management of the Bank provide access to use this database to predict credit card defaults using machine learning approaches. Therefore, this study is feasible and beneficial to the researcher as well as Bank.

Finding of the Study will help to understand possibilities of transferring to non-performing section to Credit Cards at the initial stage of the Credit Card application process. In addition, it will help to monitor and identify the customers that can pay the Credit Card regularly and the customers that not pay the Credit Card regularly.

The research findings will have a national validity of understanding the factors that affecting the Credit Cards transfer to non-performing sector and it will help to reduce non-performing Credit Cards in Sri Lankan banks.

1.8 Structure of the Study

This Study is organized as follows

- **Chapter One – Introduction**

  This chapter provides the framework of the overall research study, which is to predicting probability of Credit Card default at the stage of Credit Card application stage using supervised learning approaches. It contained a brief description of the background and scope of the research followed by the study of objective and literature review of the study.

- **Chapter Two – Literature Review**

  This Chapter provides a high-level overview of concepts of loans and advances of banks, non-performing loans and advances classification, credit scoring modes and supervised learning, un supervised learning and semi supervised learning is the introduced and discussed. The implementation of details of both supervised and unsupervised techniques are described.

- **Chapter Three - Methodology**

  This Chapter discussed about methodology used to analyze the data set and used techniques for the identify the prediction. In addition to that describe about training data set and test data set and attributes in relating to selected data set.

- **Chapter Four – Analysis and results**

  This chapter evaluates the outcomes of the implemented machine learning techniques and proposes a framework that can be used to predict the credit card defaults.

- **Chapter Five – Conclusions and recommendations**

  In this chapter summarizes the key contributions of this project and highlights opportunities for future research in addition to recommendations and conclusions.

Chapter Two

Literature Review

2.1 Introduction

Banking system is the life blood of the economy. Without a proper banking system, economic stability and growth cannot be achieved. Banks get deposits from depositors and grant loans and advances by using those deposits, to the general public. Those loans and advances are keys for the growth of investment of the country and are important parts of any organization of the country (U. Aslam, T.I.H. Aziz, A. Sohail, K.N. Batcha, 2019)These loans and advances are being given to individuals as well as corporates. Increasing competition within the financial industry, added more value to the loans and advances portfolio of the financial institutions, however granting loans and advances will increase the risk of the financial institutions. According to Investopedia, credit risk is the risk or possibility of arising loss, as a result of non-repayment of loans & advances and breach the obligations. As a result of credit risk, lender do not receive principal amount or interest as agreed. Loans & advances portfolio of financial institutions consists of term loans, credit cards, pawning, leasing and other loans. Credit card portfolio is a most important part of the Banks' lending portfolio. In the bank's loans & advances portfolio, credit card has major portion and importance.

2.1.1 Non-Performing Loans (NPLs)

When considering about Non-Performing Loans (this include credit card NPL), there is no standard definition for the same, whereas various definitions used by various organizations and institutions as well as individuals around the globe. These NPL definitions and clarifications common to all Loans and advances including credit card portfolio of the bank.

According to Hou & Dickinson (Y. Hou and D. Dickinson, 2008) Non-Performing Loan is which the full payment in principle or the interest is not paying in full in relating to granted loans and issued credit cards and these kinds of loans and credit cards are not earning income to the Banks.

According to the International Monetary Fund (IMF) (2005) NPL defined as "A Loan is non-performing when payments of interest and principal amount are past due for 90 days or more, or at least 90 days of interest payments have been capitalized, refinanced or delayed by agreement, or payments is less than 90 days overdue, but there are other good reasons to doubt that payments will be made in full"

Non-Performing Loans & Advances is the main issue of the modern banking. The major attention paid by the researchers to this area has been increased in last two decades. Development of new accounting standards/compliance standards and minimization of risk weights of NPL is much needed for Banks and financial institutions. When considering about developing countries, NPL is increasing in a higher rate (O.M. Faruk, S.M. Islam, n.d.). In Sri Lankan Banks, NPL ratio is in an average level of 5% in last five years. Increasing trend of Non-Performing Loans and advances is directly hit to the banks' balance sheet as well as the income statement. Provisions for NPLs are directly affect to the Bank profitability. In addition, increasing NPLs directly affect to the refinancing ability of the Banks and reduced the availability of loanable funds. Most of the Banks failed due to higher level of Non-Performing Loans and advances. Specially in 2007 -2008 during subprime mortgage crisis the Bank's NPLs increased in hyper rate. The managements of those Banks used various strategies to reduce NPL ratio of the organizations. However, it was difficult to maintain it in precise level.

According to Bank of International Settlement (BIS) Loans and advances classify into five main categories.

- Passed
- Special Mention
- Substandard
- Doubtful
- Loss

According to Bank of Ceylon classification, Loans & Advances and Credit Cards classified into two main categories

- Performing Loans/ Credit Cards
- Non-Performing Loans/ Credit Cards

Non preforming loans / Credit Cards section re-classified as

- Special Mention
- Substandard

- Doubtful
- Loss

2.1.2 Non-Performing Credit Cards

Non-Performing Credit Cards is one sub category of the Non-Performing Loans portfolio. Other than the Credit Cards issued under fully secured category, all the other Credit Cards are also classified under Performing and Non- Performing, if the minimum payment of credit card arrears for ninety (90) days or more, less hundred and twenty-one (120) days from due date is classified under “Special Mention” category. Above hundred and twenty (120) days or more and less hundred and eighty-one (181) days is classified as “Substandard” and above hundred and eighty (180) days or more and less than two hundred and forty-one (241) days is classified as “Doubtful” and two hundred and forty (240) days or more is classified as “Loss”.

When considering about other loans and advances the type of these classifications are different. Different months and dates are taken into consideration for term loans, housing loans, personal loans etc.

To minimize the credit risk and increase the credit worthiness and quality of the Credit Card portfolio and other loans portfolio the Banks use credit scoring methods. Most of the credit scoring methods calculated using manual methods in banking context. Specially, The Bank of Ceylon calculates the credit scoring using various manual methods for different categories of loans & advances. For corporate and retail sector different credit scoring method and for credit cards different credit scoring methods are being used. All of these methods maintain manually and there are lot of human interventions for this credit scoring process.

2.2 Credit Scoring

Almost all the Banks and financial institutions are developed strategies to reduce Credit Risk. For that they developed credit evaluation methods and by using that evaluation methods they get crucial credit management decisions (R.A. Itoo, A. Selvarasu, A.J. Filipe, 2015). Credit evaluation methods are collected, analyze and classify different items in relating to credit and through that the credit decisions will obtain. Quality of the loans is the key determinant of profitability and survival of the Banks and financial institutions.

Competitiveness of the financial industry is increasing day by day. As a result of that necessity for a Credit Scoring model has also increased (U. Aslam, T.I.H. Aziz, A. Sohail, K.N. Batcha, 2019). Researchers have done various researches in relating to Credit Scoring and there are several methods that can be used to score the credit. Both judgmental (deductive) and statistical credit scoring models use to score the credit worthiness of the customer (R.A. Itoo, A. Selvarasu, A.J. Filipe, 2015).

Credit Scoring can define as "The process of modeling creditworthiness by financial institutions" (D.J. Hand, S.D. Jacka, 1998). According to Itoo et al and Abdou et al (R.A. Itoo, A. Selvarasu, A.J. Filipe, 2015) (H. Abdou and J. Pointon, 2011) Credit Scoring is "Set of decision models and their underlying techniques that aid lenders to granting of consumer credit. These techniques decide who will enhance the profitability of the borrowers to the lenders". Without using credit scoring models and techniques it is unable to issue and monitor the credit portfolio of the Bank.

2.2.1 Credit Scoring Models

Researchers done many studies in relating credit scoring models and some studies are listed below.

According to Itoo et al and Hand et al (R.A. Itoo, A. Selvarasu, A.J. Filipe, 2015) (D.J. Hand, S.D. Jacka, 1998) used Latent- Variable model to measure the aspects of the behavior of credit customer. This model divides the characteristics into two major variables called primary and behavioral characteristics. After analyzing the characteristics in both variables, it summarizes them and measure the overall credit score.

Xiao-Lin Li and Zhong (X.L. Li, Y. Zhong, 2012) introduced model for credit scoring called "Ensemble Learning Model". This model combines the static soring to behavioral scoring and maximize revenue by minimizing the Type 1 and Type II errors. But in this model, some information of the applicants can be inaccurate for a specific degree and have missing values.

According to Bellotti and J Crook (A. Bellotti, J. Crook, 2009) introduced a credit scoring model that includes macro-economic variables such as unemployment rate, interest rate. This model gives more accurate and significant results.

Azam et al (R. Azam, M. Danish, S. Akbar, 2012) studied the importance of loan applicant attributes in relating to socioeconomic attributes on personal loan decision, using the descriptive statistical methods and logistic regression. Based on the result arrived out of six variables, three variables have significant impact on the loan decision. Those variables are region, residence and number of years working in the current organization.

Matthew and Boateng (N.G. Matthew and S.S Boateng, 2013) conducted investigation in relating to credit and risk associated with banking sector of Ghana. It was identified that the Banks used CAMPARI for credit scoring (Character, Ability, Model, Purpose, Amount, Repayment and Insurance).

Koh et al (H.C. Koh, W.C. Tan, C.P. Goh, 2006) identified that data mining techniques can be used for credit scoring and constructed a model with five steps called defining the objectives, selecting variables, selecting sample, collecting data, selecting modelling tools and constructing the model and finally validating and assessing.

Using above credit scoring models', the Banks analyses the credit worthiness of the customers using manual and automated systems. Statistical techniques and advanced techniques are used to develop the credit scoring models. Specially both supervised and unsupervised machine learning methods are used to predict the credit worthiness of the customer, using various statistical models and various attributes.

2.3 Supervised and Unsupervised Learning

Machine learning can divide into two main branches called Supervised Learning and Unsupervised Learning (A. Goyal, R. Kaur, 2016). Supervised Learning again divided into two branches called as Classification and Regression and unsupervised learning divided into two branches called Clustering and Dimensionality Reduction.

In supervised learning dataset includes with features and labels and in unsupervised learning dataset has no labels (A. Goyal, R. Kaur, 2016).

2.4 Default prediction models

Recent Researchers have paid more attention to apply machine learning algorithms and neural networks for credit scoring and risk assessments of the Banks. These techniques consist with both traditional and advanced statistical tools and techniques (U. Aslam, T.I.H. Aziz, A. Sohail, K.N. Batcha, 2019) (Y. Hou and D. Dickinson, 2008). In addition to these classifications these techniques can divide into three main categories called statistical techniques, classical machine learning techniques and ensemble classifiers (S. Neema and B. Soibam, 2017).

2.4.1 Statistical Techniques used for credit default prediction

2.4.1.1 Logistic Regression

Using maximum likelihood method estimates the probability of an event in binary regression model However, in logistic regression model is used cumulative density function that is in sigmoid nature and estimate probability of occurrence of each event. Jayadev et al and Lee et al (M.Jayadev, N.M. Shah, R. Vadlamani, 2019) (T.S. Lee, C.C. Chiu, Y.C. Chou, C.J. Lu, 2006) used logistic regression for credit scoring model for personal and credit card loans. Easy implementation, sturdy performance and simple understanding is the main advantages of the logistic regression (U. Aslam, T.I.H. Aziz, A. Sohail, K.N. Batcha, 2019) (C.J Nali´ and A. Švraka, 2018) in normal regression the output is given as a negative value or greater than one value but in logistic regression it provides the continues range of grades between 0 to 1 and keeping output between 0 to 1.

Aslam et al and Baesens et al (U. Aslam, T.I.H. Aziz, A. Sohail, K.N. Batcha, 2019) (B. Baesens, D. Roesch, H. Scheule, 2016) developed a model using logistic regression for " Korean Student Aid Foundation" that consists of 127,432 loans and this set consists of 2,141 Non Performing Loans for training set and 83,560 loans which consists of 1,480 Non Performing Loans for test set and find that loans are defaulting due to factors named age, house hold income, field of study and monthly income. This model has 69.7% accuracy for the test data set. Similar type of study done by Agbemava et al (E. Agbemava, I.K. Nyarko, T.C. Adade, A.K. Bediako, 2016) used by logistic regression to search loan defaulters in Ghana   and they found that logistic regression-based model predicts loan defaulters perfectly with an accuracy of 86% (E. Agbemava, I.K. Nyarko, T.C. Adade, A.K. Bediako, 2016).

Goyal and Kaur (A. Goyal, R. Kaur, 2016) used logistic regression to build model tree that consists of a data set of 13 attributes and the result accuracy is between 69 % to 80% in five runs. For this test they have used R programming language.

Gultekin and Sakar (B. Gultekin and E.B. Sakar, 2018) used logistic regression to predict the loan defaults by using 16,000 data set with 18 attributes. This data set consist of 11,000 good loans and 5,000 bad loans. All variables significant in the p_value of 0.05 level according to logistic regression results. This happens due to imbalanced data set used by the researcher and also data set not uniformly distributed. After run the model classification, the performance shows that 81.2% accuracy, 78.3 % TPR and 73.4% AUC in relating to the logistic regression.

Neema and Soibam (S. Neema and B. Soibam, 2017) used machine learning methods to achieve most cost-effective prediction for credit card default using seven machine learning methods. 30,000 data set used for analyze and 23 attributes included in that data set. Using different cost factors analyze the seven methods and for cost factor = 10 logistic regressions show higher cost with 0.26 MCC and cost factor = 15 show higher cost 14,721 and 0.26 MCC (Matthew's Correlation coefficient). this study shows that credit card default depends non linearly on various factors.

Tudor et al (L.A. Tudor, A. Bara, V.S. Opera, 2017) compare the data mining methods to predict the probability of credit default in banking sector using financial data in Romania banks. Data set consists of 18,239 instances and 1,489 has recorded as NPL. They have used Logistic regression, Naïve bayes and Support vector machines for model and find that Logistic Regression perform very well in predicting defaulters.

Hassan and Mirza (M.M. Hassan and T. Mirza, 2020) develop a model and analysis Credit card default prediction using Artificial neural Networks. In this model researcher used Logistic Regression for analyze the data and indicate that only certain factors are more significant.

Torvekar and Game (N. Torvekar and S.P. Game, 2019) develop predictive analysis of credit score for credit card defaulters using four machine learning techniques and find that accuracy of 80.83% in Weka tool and 81.7% in KNIME tool for Logistic regression.

Results of Logistic regression varies in situation to situation and data set to data set.

2.4.1.2 Naïve Bayes Algorithm

Naïve bayes algorithm is based on the well-known concept on probability that called Bayes theorem. Most of the researchers used Naive Bayes for classification.

Tudor et al (L.A. Tudor, A. Bara, V.S. Opera, 2017) used Naïve bayes algorithm to predict credit defaults in Romanian banks. this algorithm predicts 95.81% correctly out of 16,750 instances with value 0 and 99.93% out of 1,489 NPA cases. But in this study researcher revealed that logistic regression is more accurate and perform well than the Naïve bayes algorithm.

Neema and Soibam (S. Neema and B. Soibam, 2017) used seven machine learning methods to analyze credit default pattern using 30,000 credit card data to identify best cost-effective prediction model in different cost factors this algorithm shows medium level of cost comparing the other Six machine learning methods.

Chou and Lo (T. Chou and M. Lo, 2018) compare the credit card defaults using deep learning and other machine learning models for this comparison used Naïve bayes as one of the conventional machine learning technique. Researcher find that Naïve bayes algorithm has 90.9% accuracy, 99.8% precision, 89.4% Recall and 0.943 F- Score. But they concluded that rather than conventional machine learning techniques Deep learning performed better.

Hamid and Ahmed (J.A Hamid and M.T. Ahmed, 2016) developed a prediction model of credit risk of banks using data mining methods and use Naïve bayes. Weka tool used for analyzing this data set and data set consist of eight attributes consisting of both nominal and numerical data. In this model Naïve Bayes correctly classified instances with 73.87% but accuracy of other methods used in this study has more accuracy than Naïve bayes.

Islam et al (S.R. Islam, W. Eberle, S.K. Ghafoor, 2019) used combined machine learning approaches to predict credit default. For this study they used publicly available data set in UCI. "Taiwan" credit card data set that has 23 features and 30,000 instances. Out of these 30,000 instances 6,626 are default/ NPL. After running the algorithm researcher identified that Naïve bayes has 67.23% accuracy, 31.82 % precision, 42.13% Recall and 0.3626 F- Score.

Torvekar and Game (N. Torvekar and S.P. Game, 2019) develop predictive analysis of credit score for credit card defaulters using four machine learning techniques and find that accuracy of 62.42% in Weka tool and 76.6% in KNIME tool for Naïve Bayes.

2.4.2 Classical Machine Learning Techniques used for credit default prediction

When considering classical machine learning techniques Neural Networks, Random Forest, K Nearest Neighbor, Support Vector Machines and Decision Trees are the most successful techniques that given better results (M.Jayadev, N.M. Shah, R. Vadlamani, 2019).

2.4.2.1 Decision Trees

Rooted Tree is produced by the Decision Tree method. This Decision Tree consists of Roots and Nodes and apply rule based inductive reasoning. All Decision rules obtain by navigating from the root of the tree up to the leaf, as per outcome of the test along path of the tree. This model most suitable for credit risk modeling and used by most of researchers (M.Jayadev, N.M. Shah, R. Vadlamani, 2019) (T.S. Lee, C.C. Chiu, Y.C. Chou, C.J. Lu, 2006).

Tejaswini et al (J. Tejaswini, M.T. Kavya, N.D.R. Ramya, S.P. Triveni, R.V. Maddumala, 2020) used Decision tree (C5.0) for accurate loan approval prediction based on machine learning approach. in this study used basic decision tree algorithm According to Patibandla and Lakshmi (R. Patibandla et al, 2017) that requires all features should be discretized and feature selection is based on highest information gain of feature set. Other than this Decision tree model they used several other machine learning methods.

Madane and Nanda (N. Madane and S. Nanda, 2019) analyze the loan prediction using Decision Tree approach using R Studio. Data set consists of seventeen different attributes in relating to credit portfolio of the banks. Using this method made the model to approve or reject the loan application. Researcher find that credit history applications that do not pass the guidelines are mostly not approved and low-income applicants are more likely to receive approval because most of low-income applicants are more likely to repay the loans.

Supriya et al (P. Supriya, M. Pavani, N. Saisushma, V.N. Kumari, K. Vikas, 2019) develop loan prediction by using machine learning models they used several machine learning techniques such as support vector machines, K Nearest Neighbor, Gradient Boosting and Decision Tree to analyze data set consists of both qualitative and categorical data with 12 attributes and reveal that Decision Tree has the highest accuracy of 81.1% when comparing other techniques. Decision tree given the advantage of interpretability.

Batura et al (F. Butaru, Q.Chen, B. Clark, S. Das, W. Andrew, 2016) has developed model for analyze risk and risk management in the credit card industry using Decision Tree (C4.5), Logistic Regression and Random Forest. For this study used data in relating to six major U.S financial Institutions. Researcher find that Decision Tree and Random Forest outperform than the logistic regression.

Neema and Soibam (S. Neema and B. Soibam, 2017) compared the machine learning methods to achieve most cost-effective prediction for credit card default and find that Decision tree perform well and has accuracy more than 80%.

Agbemava et al (E. Agbemava, I.K. Nyarko, T.C. Adade, A.K. Bediako, 2016) predict Credit Card defaults with deep learning and other machine learning models and find that Decision Tree has accuracy of 76.2%, precision 76.2%, recall 100% and F-Score 0.865 and perform well in the prediction.

2.4.2.2 K – Nearest Neighbour

KNN is well known instance – based clustering model and this method also used for predicting credit defaults.

Goya and Kaur (A. Goyal, R. Kaur, 2016) compared the machine learning methods to achieve most cost-effective prediction for credit card default and find that K- Nearest Neighbour perform well and has accuracy more than 70%.

Chou and Lo (T. Chou and M. Lo, 2018) predict Credit Card defaults with deep learning and other machine learning models and find that K- Nearest Neighbour has accuracy of 80.6%, precision 84.3%, recall 91.6% and F-Score 0.878 and perform well in the prediction.

Islam et al (S.R. Islam, W. Eberle, S.K. Ghafoor, 2019) develop a credit default mining using the combined machine learning and heuristic approach and find that K- Nearest Neighbour has accuracy of 82.76%, Precision of 71.34%, Recall of 36.87% and F – Score 0.4862. for this model researcher used the Credit card data set publicly available in UCI repository name “Taiwan” credit card data set that include 30,000 instances and 23 attributes.

2.4.2.3 Artificial Neural Networks (ANN)

According to Jayadev et al (M.Jayadev, N.M. Shah, R. Vadlamani, 2019) “Artificial Neural Network uses a dense network of simple nodes called neurons organized in layers linked by weighted connections to transform inputs into outputs using a non-linear activation function, typically a sigmoid or a hyperbolic tangent”.

Tejaswini et al (J. Tejaswini, M.T. Kavya, N.D.R. Ramya, S.P. Triveni, R.V. Maddumala, 2020) used Neural networks for predict accurate loan approval and get positive results when comparing other machine learning techniques.

Mbuvha et al (R. Mbuvha, I. Boulkaibet, T. Marwala, 2019) develop model using neural networks for credit card default modeling. For this model used the data set consists of 30,000 instances and 23 attributes. And developed and compared two approaches for Bayesian inference in neural networks. Both models are critically allowed interpretation of the relative feature influences on the probability of default.

Goyal and Kaur (A. Goyal, R. Kaur, 2016)develop loan risk accuracy prediction model using eleven machine learning methods and reveal that Neural network has accuracy between 75% to 83% in five runs of the model.

Hassan and Mirza (M.M. Hassan and T. Mirza, 2020) develop credit card default prediction using artificial neural networks. Researcher used data set of 30,000 customers that holding credit cards and used 18,000 for training (60%) and 12,000 (40%) for test data. There are 6,636 default customers (22%). 24 variables included in the data set. After develop the model performance indicators determine the accuracy of 79% and RMSE of 0.37.

2.4.2.4 Support Vector Machine (SVM)

According to Jayadev et al (M.Jayadev, N.M. Shah, R. Vadlamani, 2019) “SVM classifies observations into classes by creating a hyperplane in the feature space such that the distance from the hyperplane to the data points is maximized which is essentially a quadratic optimization problem and is based on the structural risk minimization principle”. Yang (Yang, 2007) create an adaptive scoring system using SVM.

Goyal and Kaur (A. Goyal, R. Kaur, 2016) develop loan risk accuracy prediction model using eleven machine learning methods and reveal that Support Vector Machine has accuracy between 76% to 81% in five runs of the model.

2.4.3 Ensemble classifiers used for credit default prediction

Some models are weak in relating to other models. In ensemble classifiers grouped the weak classifiers and build powerful model with higher classification accuracy. Jayadev et al (M.Jayadev, N.M. Shah, R. Vadlamani, 2019) following ensemble classifiers are used for predict credit default in banks.

2.4.3.1 Random Forest Algorithm

Combination of Decision trees called as Random Forest algorithm (L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, 1984) (M.Jayadev, N.M. Shah, R. Vadlamani, 2019) this collection of decision trees individually classifies an observation.

Tejaswini et al (J. Tejaswini, M.T. Kavya, N.D.R. Ramya, S.P. Triveni, R.V. Maddumala, 2020) used Random Forest for predict accurate loan approval and get positive results when comparing other machine learning techniques.

Setiawan et al (N. Setiawan, Suharjito, Diana, 2019) compare prediction methods for credit default on peer-to-peer lending using machine learning. In this comparison they identified that best accuracy was obtained by random forest with 88.5% accuracy. Researcher used 60,000 records of data for this study in relating to lending club. But in this study decision tree obtained the best precision of 97.1%. another research has done using Random Forest and obtain 78% of accuracy. For this study used the data set with 13 features and using the same data set with 4 features obtained the accuracy of 69.8%.

Butaru et al (F. Butaru, Q.Chen, B. Clark, S. Das, W. Andrew, 2016) analyze the credit default prediction using machine learning techniques. And revealed that random forest and multilayer perception has more accuracy, TPR and AUC than the Logistic regression. For this study used 16,000 instances of data and from this 16,000 5,000 were NPL. Random forest Shows the Accuracy of 84.2%, TPR of 78.8% and AUC of 82.3%.

Gultekin and Sakar (B. Gultekin and E.B. Sakar, 2018) analyze the risk and risk management in the credit card industry of six major financial institutions of the USA using several machine learning techniques and they identified that Random Forest and decision tree is perform well in predicting the risk with higher precision and accuracy.

Islam et al (S.R. Islam, W. Eberle, S.K. Ghafoor, 2019) predict the default credit card using combine approaches of machine learning and they have identified that Random

Forest perform well by showing accuracy of 94.46%, precision of 94.78%, Recall of 79.32% and F-Score of 0.8637. when comparing to other machine learning and heuristic approaches.

Goyal and Kaur (A. Goyal, R. Kaur, 2016) tested accuracy prediction for loan risk using machine learning models and identified that Random Forest has better performance with accuracy between 77% to 83% in five runs of the model.

Neema and Soibam (S. Neema and B. Soibam, 2017) has conducted the comparison of machine learning methods to achieve most cost-effective prediction for credit card default. And identified that Random Forest has the best outcome that for cost factor = 10 cost of 9,478 and in cost factor = 15 cost of 11,435 by showing minimum cost comparison to the other nine machine learning methods.

In addition to Random Forest Algorithm there are some other ensemble methods can used in predicting the default rate in credit cards such as Adaptive Boosting (Ada Boost) and Extreme Gradient Boosting (XG Boost) (M.Jayadev, N.M. Shah, R. Vadlamani, 2019).

2.5 Attributes Used in Credit Default Prediction

Researches used various attributes to develop machine learning models in relating to Credit card default prediction.

Goyal and Kaur (A. Goyal, R. Kaur, 2016) used Loan ID, Gender, Marital Status, Number of Dependents, Education Level, Employment details, Income level, Loan amount, Credit History, Property area and Loan Status as main attributes.

Neema and Soibam (S. Neema and B. Soibam, 2017) used 23 variables to predict Credit Card defaults. Credit amount, Gender, Education, Marital Status, Age, History of past Payments, Amount of Bill statement, Amount of Previous payments taken as main attributes and sub attributes including under the main attributes.

Hamid and Ahmed (J.A Hamid and M.T. Ahmed, 2016) used seven attributes to predict risk of loans. Credit history, Purpose, Gender, Credit amount, Age, Housing, Job and Class is the seven attributes that taken into consideration.

Gultekin and Sakar (B. Gultekin and E.B. Sakar, 2018) used eighteen attributes to predict default credits. Housing maturity, marital status, occupation, educational status, vehicle maturity, consumer maturity, productNum, working time, workplace, ownership code, age, insurance, class, Loan type, Credit Reporting Agency and default number are used as attributes.

Supriya et al (P. Supriya, M. Pavani, N. Saisushma, V.N. Kumari, K. Vikas, 2019) used 12 attributes for loan prediction. Gender, Marital Status, Dependents, education, employment, income, amount of loan, credit history taken as major attributes.

Tejaswini et al (J. Tejaswini, M.T. Kavya, N.D.R. Ramya, S.P. Triveni, R.V. Maddumala, 2020) used 13 attributes to develop accurate credit prediction model using machine learning. Loan ID, Gender, marital status, dependents, education, employment, income credit history, property area, loans status used as main attributes.

Most of studies has divided the data set as 70% of training data and 30% of test data. Some studies divided the data set as 60% training data and 40% of test data.

2.6 Research Gap and Conclusion

This chapter discussed about previous studies done by the researchers in relating to predict default credit facilities using machine learning approaches. Most of them used statistical techniques such as Logistic regression and Naïve Bayes and classical machine learning techniques such as Decision Trees, K – Nearest Neghbour, Artificial Neural Network and support Vector Machine in addition researchers used Ensemble classifiers such as Random Forest and adaptive boosting. According to behavior of the data set accuracy, precision and recall are different from study to study. some studies said Logistic regression is more accurate and some studies said naïve bayes accurate also some studies said other methods are more accurate and outperform. Most of studies discussed above said that ensemble method random forest outperforms than the other methods.

Above studies discussed using publicly available data for develop models and few studies used the real-life data. When considering about data sets most of data sets includes transaction details relating attributes. Also, most of demographic relating attributes also including in the data sets. Almost all of these studies are done after granting a loan or issuing a credit card. Before granting a loan or issue a credit card they do not predict the possibility of non-Performing that facility. Most of banks and financial institutions have done credit scoring methods manually and it has personal bias factors. As a result of that accuracy of that credit scorings are very low. Therefore, machine learning based credit scoring model embedded with accurate NPL prediction model is mush needed for banking and finance sector.

Also, there are few Studies in relating to develop models in the credit card application process default prediction.

Chapter Three

Methodology

3.1 Chapter Introduction

This chapter identifies how the study was done and its aim is to describe the research strategy and methods applied in this study, and to discuss their suitability within the context of various research philosophies, models and methodological approaches. This includes a general overview of the overall research philosophy employed in carrying out the research, justification of the chosen approach, provision of operational construct definitions and specification of their indicators, and a discussion of the data collection and analysis methods. It is useful to state at this point that, due to the confirmatory nature of the research objectives, the questions that emerged in chapter two and previous research foundations reported in the literature.

3.2 Steps used in Study

This study mainly focusses on identifying and predicting credit card defaulters using machine learning algorithms. Specially using the classification techniques. Figure 1 shows the major steps that are used to identify and predict credit card defaulters using supervised machine learning approaches.

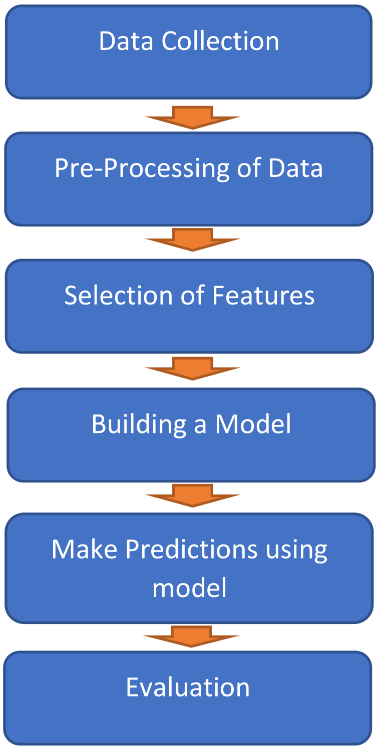


Figure 1: Steps using to predict credit card defaulters

3.2.1 Data Collection

As per figure 1 process starts from data collection. Data set that used in this study was obtained from data warehouse of Bank of Ceylon. This data set includes all data in relating to credit card users of Bank of Ceylon that apply for Credit Cards between 2017 to 2021. This data set divided in to two parts called “Training set” and “Testing set” using 80:20 ratio and to increase accuracy of the model using cross validation technique. Accuracy of data model test using machine learning approaches called Naïve Bayes, Logistic Regression, Support Vector machines, Random Forest and K Neighbors Classifier. After that technique with the highest accuracy for both data sets were selected.

The Data set of this study consists of 6,419 samples and each represented with a feature vector of 18 variables. The variable names and types shown in Table 1. The data set belongs to the individual credit card applications in relating to Bank of Ceylon. Variables of this data set can divide in to two main categories called finance related information and personal information. Account balance at the time of applying the credit card, Income of the applicant, other obligations such as loans of the customer, investments of the applicant are considered as financial information and age, number of dependents, years of living in present address, gender etc. included in the personal information category. This financial- related information as well as personal information both important in the credit worthiness of the customer. Status of the credit card show the whether a credit card balance is gone into default or not.

Number	Variable Name	Type
1	Category of Application	Categorical
2	Status of Credit Card	Categorical
3	Is the Applicant a New Customer	Categorical
4	Is there any facility previously at any financial Institution	Categorical
5	Has applicant ever guaranteed any facility at any financial Institution	Categorical
6	Employment Type	Categorical
7	Number of Credit Cards held	Categorical

8	Years of stay in current residential address	Categorical
9	Relationship with the Bank (Satisfactory running of SA/CA Accounts)	Categorical
10	Years in current employment/business	Categorical
11	Accommodation Type	Categorical
12	CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a borrower	Categorical
13	CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a guarantor	Categorical
14	Age (In Years)	Categorical
15	Existing monthly loan installments (including facilities at any financial institutions) to Monthly Net Income	Numerical
16	Monthly Net Income or Profit (Rs.)	Numerical
17	No of family members (dependents + Applicant)	Numerical
18	Gender	Categorical

Table 1: Variables of the Credit Card Data set used in this study

3.2.2 Pre-Processing of Data

Second step of this process is pre-processing of data. This step includes several processes of handling the missing values. Some missing values are eliminating and some missing values imputed using missing value imputation methods and some variables transform into new forms especially categorical variables transform into binary values. This type of study data pre- processing is one of the critical steps and it seals with the preparation and transformation from the initial data set to the final data set. This process is the more time-consuming stage of this project. From total 6,419 samples removed fully secured customers and staff customers and taken all normal customers totaling 6,321 in to study.

Data cleaning of loan data removed several attributes that has no significance about the behavior of a customer. Data integration, data reduction and data transformation are also applicable for this credit card data set. In this analysis the data is reduced to some minimum of records. Initially the attributes which are critical to make a credibility of prediction is identified with information gain as the attribute- evaluator and ranker as the search method.

3.2.3 Feature Selection

In this step select the features and building the classification model. It predicts that the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The derived model is based on the analysis set of training data i.e., the data object whose class label is well known. Using the machine learning algorithm feature selection can be achieved and the targeted learner model can be built.

3.2.4 Make Predictions using model

In this step identify and distribute trends based on the available data set. The model is tested using the test dataset and make predictions.

3.2.5 Evaluation

In the final stage, the designed system is tested with test set and the performance is assured. Evaluation analysis refers to the description and model regularities or trends for objects whose behavior changes over time. For this evaluation commonly used the precision and accuracy matrixes. For increase the accuracy of the model used cross validation model in the event of evaluation. Figure 2 shown that detailed flowchart of model used to analyze the data set

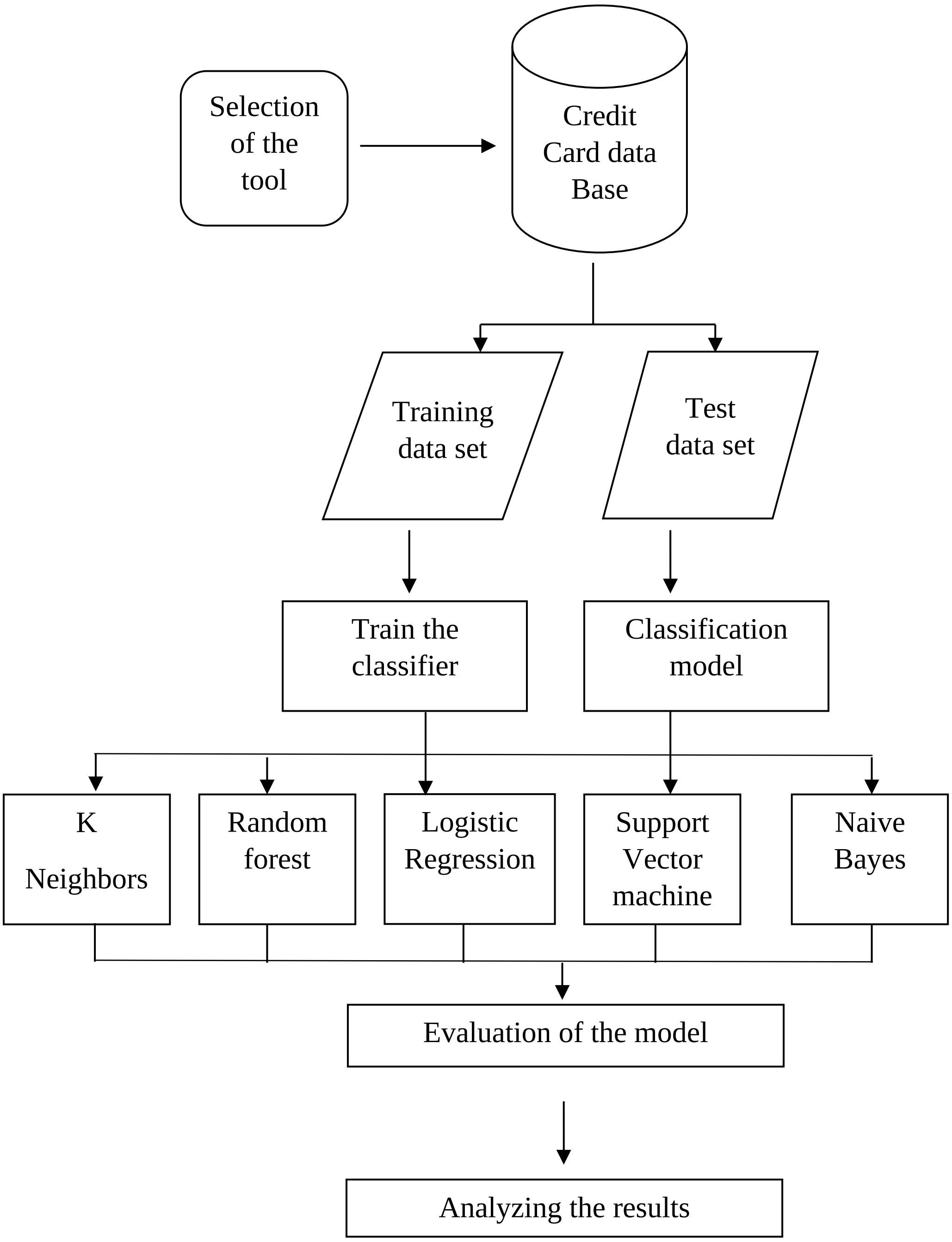


Figure 2: Proposed model of the study

3.3 Application flow of the study

Using this study, the Banks can predict and verify the eligibility of the customer for credit card using this application. The customers or staff of the Bank can check loan eligibility directly online by providing required information and the response will receive as soon as he/ she apply to the credit card whether he/she is eligible for credit card or not. Figure 2 shown the application flow of Bank credit card. After apply this application customers not required to visit bank staff for apply credit cards and it will save money and time of the individual and helps in better user experience as well as increase the efficiency of the bank by saving time and increase the accuracy of the credit scoring process of the bank. Figure 3 shows that application process graphically.

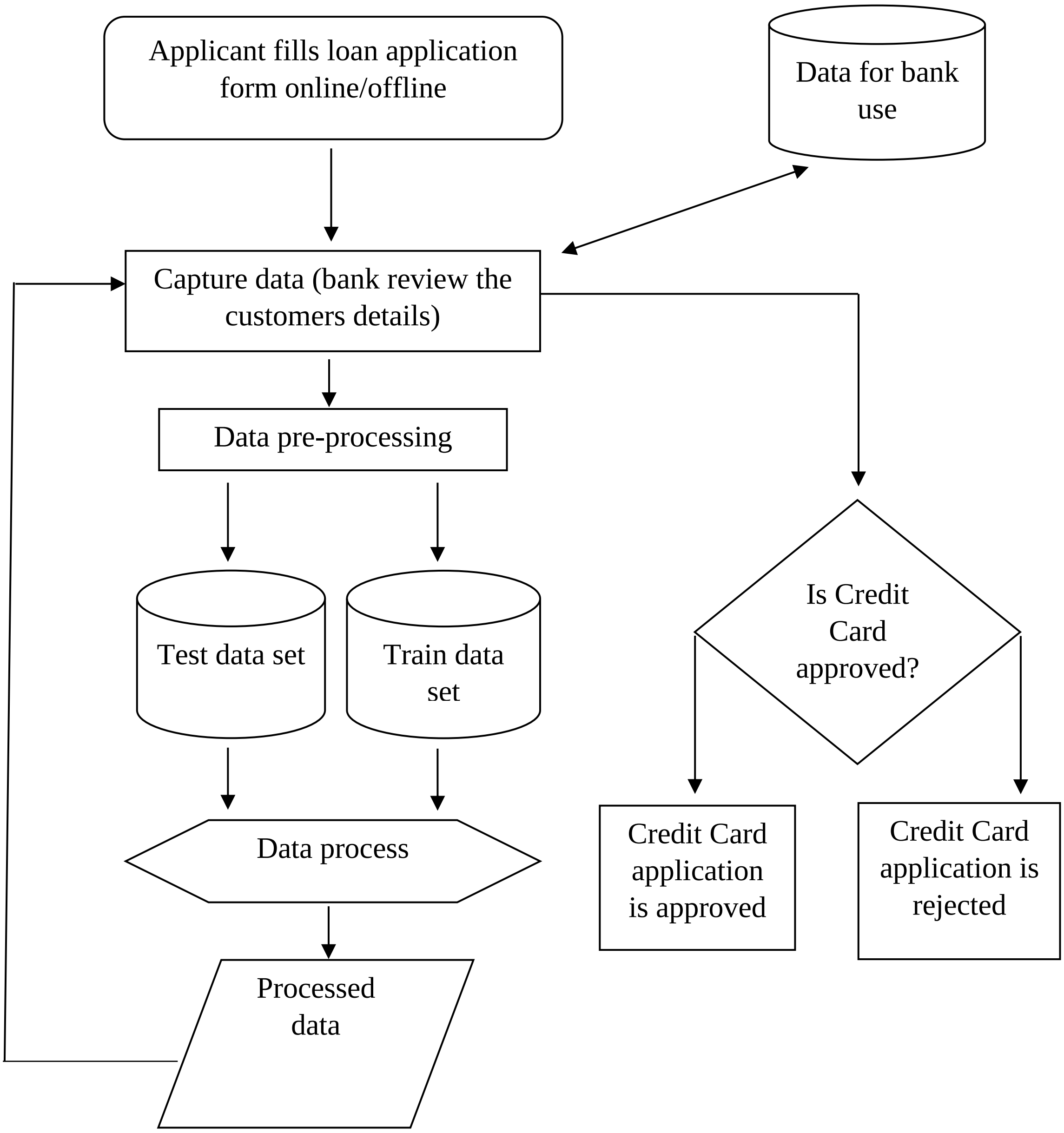

Applicant fills loan application form online/offline
Data for bank use
Capture data (bank review the customers details)
Data pre-processing
Test data set
Train data set
Is Credit Card approved?
Data process
Credit Card application is approved
Credit Card application is rejected
Processed data

Figure 3: Credit Card application flow

3.4 Role of the Researcher

Researcher expects to broaden bankers' understanding regarding the credit scoring process of the Banks and help them to issue high quality credit cards and minimize the non – performing credit cards percentage. This will improve the efficiency as well as profitability of the Bank.

Specially, researcher expects to use the findings of this study to guide Bank of Ceylon in addressing the issue of increasing non-performing loans and advances specially in credit card portfolio. Also, this study helps to automate the credit card issuing process of the Bank and remove the manual credit scoring process. In addition to that, researcher helps the Bank to take foot forward into banks digitalization of processes.

3.5 Validity

Findings of this study help to identify the possible credit card defaulters in the process of the application. As a result of that the Bank can minimize the non-performing credit cards at the stage of credit card application. This study helps Bank of Ceylon to whom to approve credit cards and whom to not approve credit cards in systematic way rather using manual credit scoring models that contain personal bias.

3.6 Reliability

A test is seen as being reliable when it can be used by a number of different researchers under stable conditions, with consistent results and the results not varying. Reliability reflects consistency and replicability over time. Furthermore, reliability is seen as the degree to which a test is free from measurement errors, since the more measurement errors occur the less reliable the test (Fraenkel & Wallen, 2003; McMillan & Schumacher, 2001, 2006; Moss, 1994; Neuman, 2003). In this study we measure … variables relating to credit card application and run the model. Those variables are stable and constant. It produces the same results if the same individuals and conditions are used therefore errors are minimized. Therefore, these measurements enhance the reliability up to greater extent.

3.7 Generalizability

This study commenced referring Bank of Ceylon and its customers. The outcome of this study can be used for understanding of the ideal customers for banking industry to approve and issue the credit cards. Instead of just done credit scoring process manually this model provides automated facility to do credit scoring it will reduce time and cost of credit scoring and can identify future credit card defaulters correctly. Therefore, this model can use in other Banks in future through their credit card data base.

3.8 Ethical Considerations

Throughout the research process the researcher faces ethical challenges in all stages of the study. The possible ethical challenges include anonymity, confidentiality, informed consent, researchers’ potential impact on the participants and vice versa as discussed by DesRoches et al., 2008.

In this research carefully investigates and understand ethical behaviors relating to the research. Always looking for that any harm arising for humans relating to all aspects of the research that develop. In this research follow all principles relating to ethical consideration according to Bryman and Bell (2007). Participants of this research not subjected to harm in any case. Privacy of respondents assured for all participants and responses of participants should not use to gain any competitive advantage. Respect for the dignity of research participants prioritized in this research. This research not proposed to process of genetic information or personal data (e.g., health, sexual life style, ethnicity, political opinion, religious or philosophical conviction). In order to secure the Banks internal data set, the researcher adopt confidentiality and the information not be exposed to outsiders.

All data relating to research collect prior to the study. All type of communication in relation to the done with honesty and transparency. All type of misleading information, as well as representation of primary data findings in a biased avoided in this research. Ethical consideration upholds according to the university guidelines, disclosing resources and using Harvard reference system.

Chapter Four

Investigation and Analysis

4.1 Introduction

This chapter will focus on presenting the data which gathered through the previous chapter of the research. The data was gathered using Bank of Ceylon credit card data base. The characteristics will be illustrated graphically. Each variable would be presented with statistical background. Also, this chapter will present a detailed analysis of the outputs and supervised machine learning algorithms used. The remainder of this chapter outlines the main research findings of this study.

4.2 Composition of the data set

The data consists of 6,419 customers and 18 variables. Each sample corresponds to a single customer. The variables consist of this data set shows in the Table 01 in methodology chapter in detail. Summary of variable are as follows.

1. Category of Application
2. Status of Credit Card
3. Is the Applicant a New Customer
4. Is there any facility previously at any financial Institution.
5. Has applicant ever guaranteed any facility at any financial Institution
6. Employment Type
7. Number of Credit Cards held
8. Years of stay in current residential address
9. Relationship with the Bank (Satisfactory running of SA/CA Accounts)
10. Years in current employment/business
11. Accommodation Type
12. CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a borrower
13. CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a guarantor
14. Age (In Years)

15. Existing monthly loan installments (including facilities at any financial institutions) to Monthly Net Income
16. Monthly Net Income or Profit (Rs.)
17. No of family members (dependents + Applicant)
18. Gender

The variables Age, Gender, Customer category, Employer type are defined as demographics variables, since they describe a demography of customers and are available for new customers.

The total proportion of defaults (NPA Customers) in the data is 8.1% which is 523 out of the total data set comprising of 6,419 remaining samples. This number of defaulters shows that realistic representation of the bank's customer base and industry norms. Therefore, generalization of results can do fairly using this data set. Above variables describe as follows.

4.2.1 Category of Application

This Variable indicates the category of applicant. There are three types of applicants called normal customers, fully secured customers and staff of financial institutions customers.

4.2.2 Status of Credit Card

This variable indicates whether or not the customer defaulted in their credit card debt payment. For the purpose of this project, predicting default is the main focus of the data analysis. A value of "1" indicates default, and a value of "0" indicates no default.

4.2.3 Is the Applicant a New Customer

Existing customers and new customers of Bank can request credit cards therefore this variable shows that the customer requested credit card is new customer or existing customer of the Bank.

4.2.4 Is there any facility previously at any financial Institution.

Customers can apply and have credit card facilities and any other credit facilities from other banks. This variable shows that applicant has previously borrowed from other financial institution.

4.2.5 Has applicant ever guaranteed any facility at any financial Institution.

Customers can sign as a guarantor for other loans and advances. This variable shows that applicant of new credit card previously signs or not as a guarantor for other loans and advances.

4.2.6 Employment Type

In this variable, employee type categorized as government employee, private sector employee, professionals and self-employed.

4.2.7 Number of Credit Cards held

In this variable customer categorized according to number of credit cards holding of customer in other financial institutions. Categorized as less than or equal to 3 cards, greater than 3 cards and no card hold.

4.2.8 Years of stay in current residential address

In this variable customer categorized according to years of stay in current residential address. Categorized as less than one year, more than or equal to one year and up to three years, more than 3 years and up to five years and more than five years.

4.2.9 Relationship with the Bank (Satisfactory running of SA/CA Accounts)

In this variable customer categorized according to relationship with bank. If customer holding and satisfactory running savings or current account are check in this variable. Customers categorized as no account with bank, not operative/ not satisfactory, account operative up to six months, account operative more than six months and up to twelve months, account operative more than twelve months and up to eighteen months, account operative more than eighteen months and up to twenty-four months, account operative more than twenty-four months.

4.2.10 Years in current employment/business

In this variable customer categorized according to years in current employment or business. Customers categorized as less than one year, more than or equal to one year and up to three years, more than three years and up to five years, more than five years.

4.2.11 Accommodation Type

In this variable categorized the customer accommodation type according to own house, rented and live with parents.

4.2.12 CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a borrower

In this variable shows that this customer is previously not paid the obligations and appearing in report of credit information bureau in Sri Lanka in last two years.

4.2.13 CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a guarantor

In this variable shows that this customer is appearing in report of credit information bureau as guarantor in Sri Lanka in last two years.

4.2.14 Age (In Years)

In this variable categorized the customer in to four categories called up to 25 years, greater than 25 and less than or equal to 35, greater than 35 and less than or equal to 45 and greater than 45 years.

4.2.15 Existing monthly loan installments (including facilities at any financial institutions) to Monthly Net Income

4.2.16 Monthly Net Income or Profit (Rs.)

4.2.17 No of family members (dependents + Applicant)

4.2.18 Gender

In this variable categorized gender as Male and Female.

4.3 Uni – and bivariate analysis

To gain better understanding of the characteristics of the dataset, a uni and bi-variate analysis comparing descriptive statistics and distributions of the individual variables was carried out.

4.3.1 Demographic Variables

In this section describe about behavior of demographic variables mainly Age, Gender, Number of Family members, employment type, accommodation type, Years of stay in current residential address and Years in current employment/business

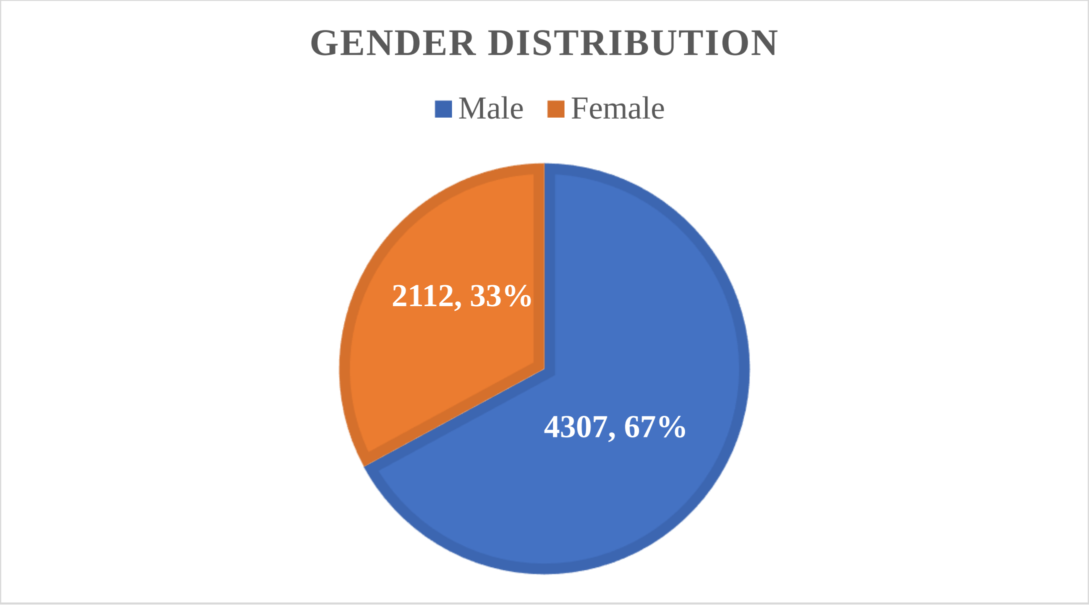


Figure 4: Gender Distribution

When considering Gender Distribution figure 4 shows that this data set represent 67% of Males and 33% of females. Therefore, by looking at this distribution can conclude Males are requesting credit cards than females.

When considering Number of family members figure 5 shows that 33.1% of applicants have 2 family members and 29.2% applicants have one member. Minimum family members are 0 and maximum family members are 8.

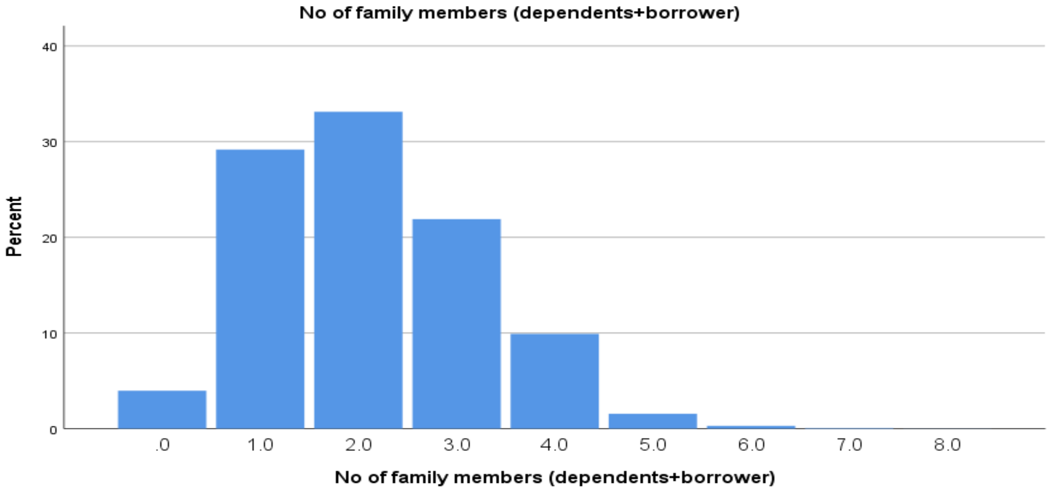


Figure 5 : Number of Family Members

According to figure 6 shows that most 47.5% of applicants represent the age category of greater than 25 and less than or equal to 35 age range and 26.5% of applicants represent the grater than 35 years and less than or equal to 45 years of age. Age below 25 and age above 45 categories represent 26.0% of total applicants.

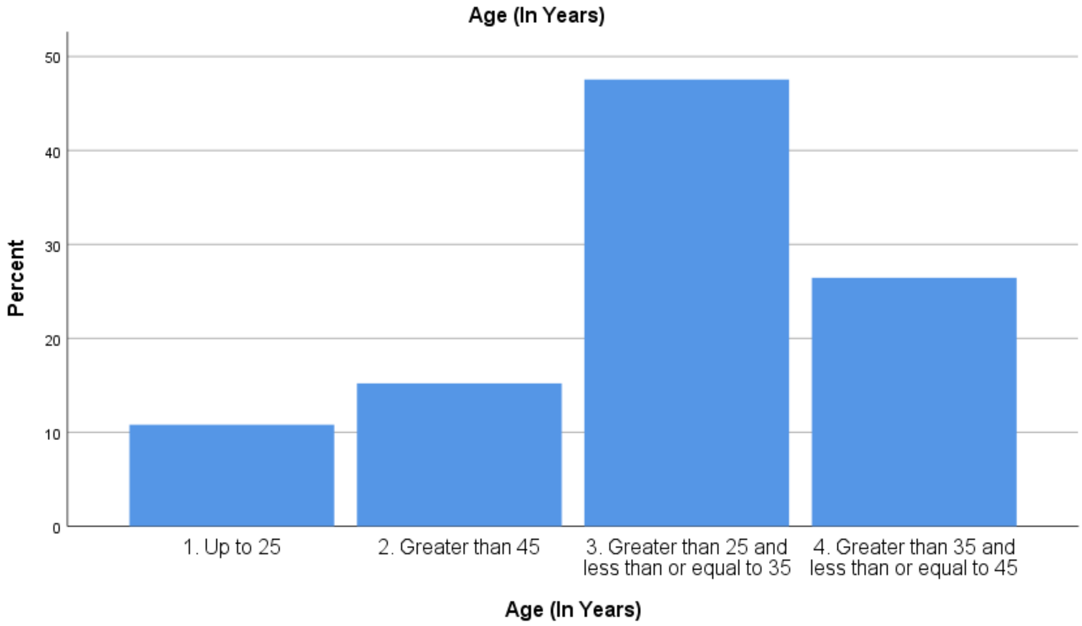


Figure 6 : Age Category

When considering employment type figure 7 shows that 76.3% of applicants represent the category of permanent employees of government and statutory bodies and all other employment categories represent the 24.7%. therefore, we can say Credit card applicants of bank of Ceylon represent government sector employees and private sector employees, professionals and businessmen not interesting to apply the credit cards offered from Bank of Ceylon.

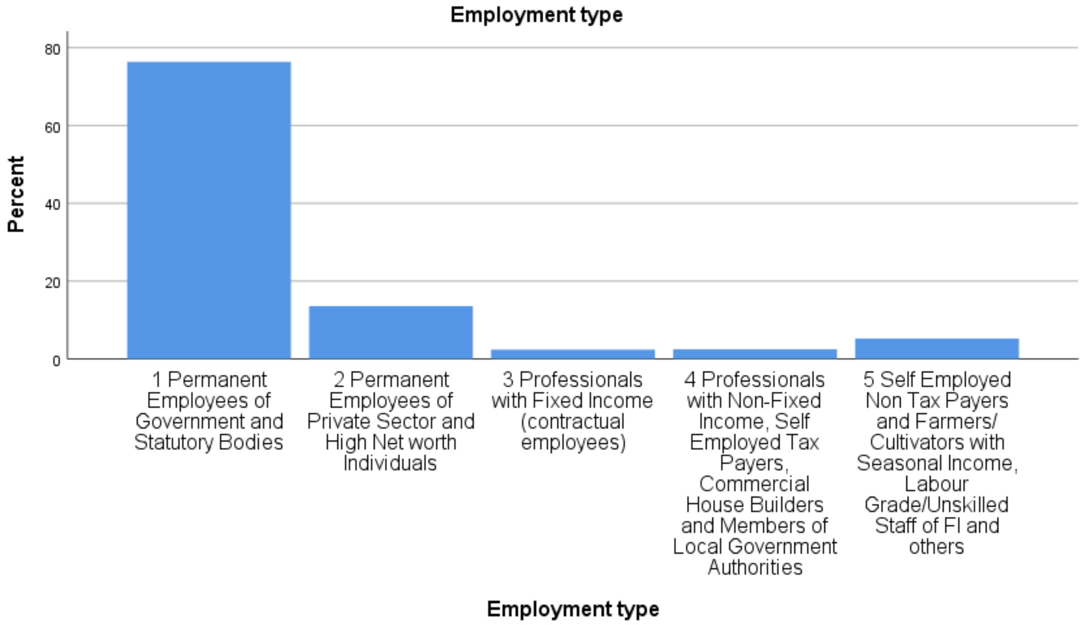


Figure 7 : Employment Type

When considering Accommodation type figure 8 shows that 72.2% of applicants lived in their own houses and 26.2% of applicants lives with houses that their parents owned. Only 1.6% of applicants lives in rented houses.

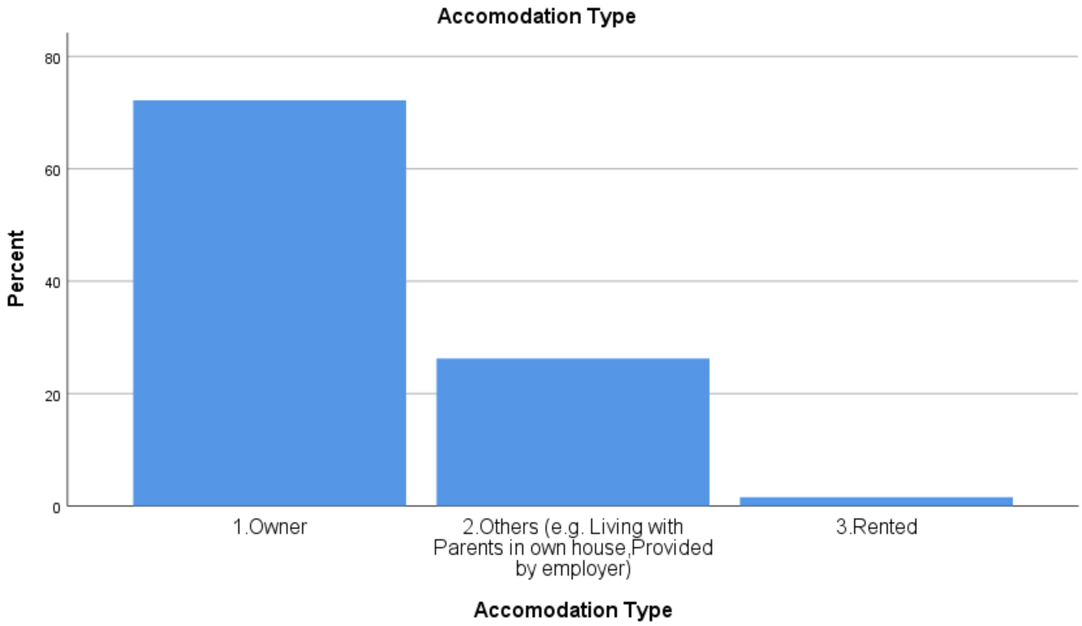


Figure 8: Accommodation Type

When considering Years in current employment figure 9 shows that 68.7% of applicants works with current employer/ business more than 5 years and only 6.1% of applicants works in current employment less than one year. Therefor we can say most applicants are well established with their current employment.

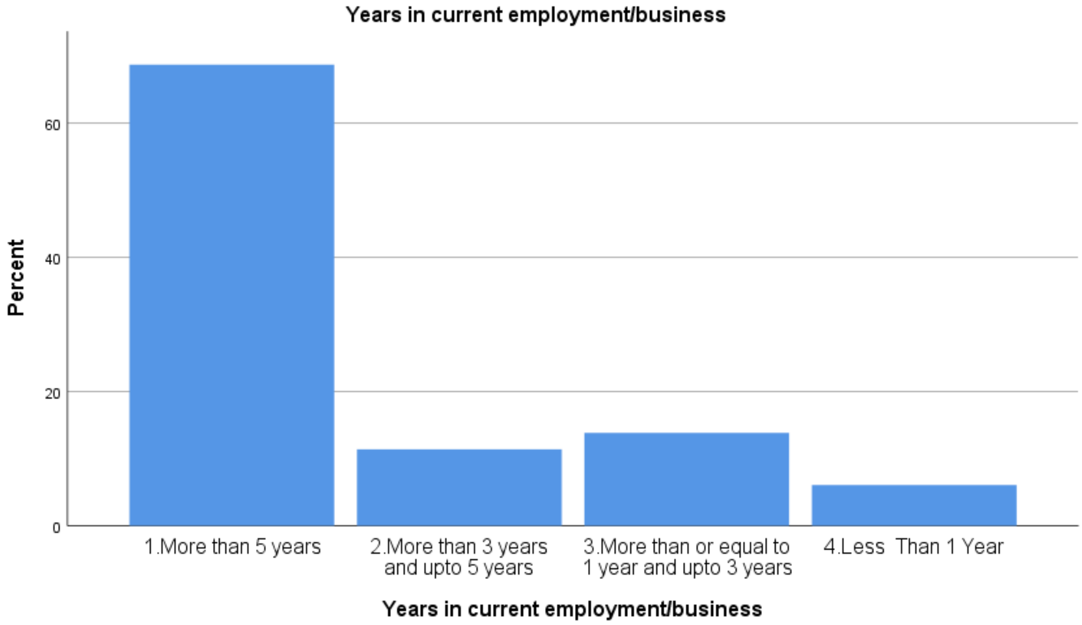


Figure 9: Years in Current Employment/ Business

Figure 10 Shows that distribution of applicants that number of years of stay in current residential address and according to figure 10 94.6% of applicants are stay in current residential address and only 0.8% of applicants lives in current residential address less than a year. This shows that applicants of credit cards of Bank of Ceylon are well established in their current residential address.

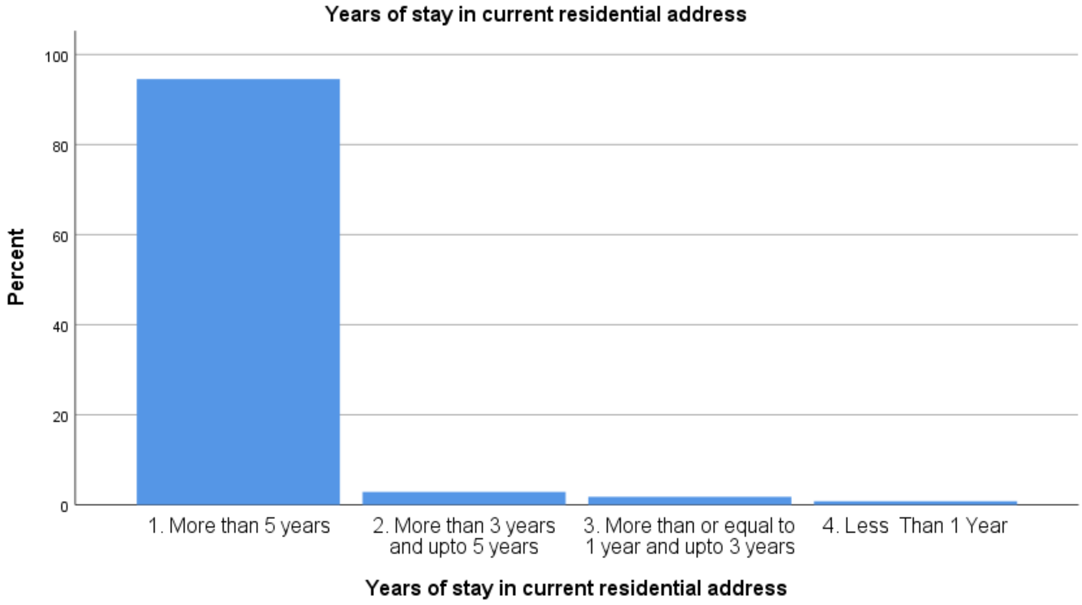


Figure 10 : Years of stay in current residential address

4.3.2 Financial Variables

In this section describe about the financial variables in relating to this study. Is the applicant a new customer, Is there any facility previously at any financial Institution, Has applicant ever guaranteed any facility at any financial Institution, Number of Credit Cards held, Relationship with the Bank (Satisfactory running of SA/CA Accounts), CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a borrower, CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a guarantor, Monthly Net Income or Profit (Rs.) are financial variables representing this study.

According to figure 11 shows that distribution of customer is new applicant or not. New applicant means the applicant does not have any relationship with bank previously.

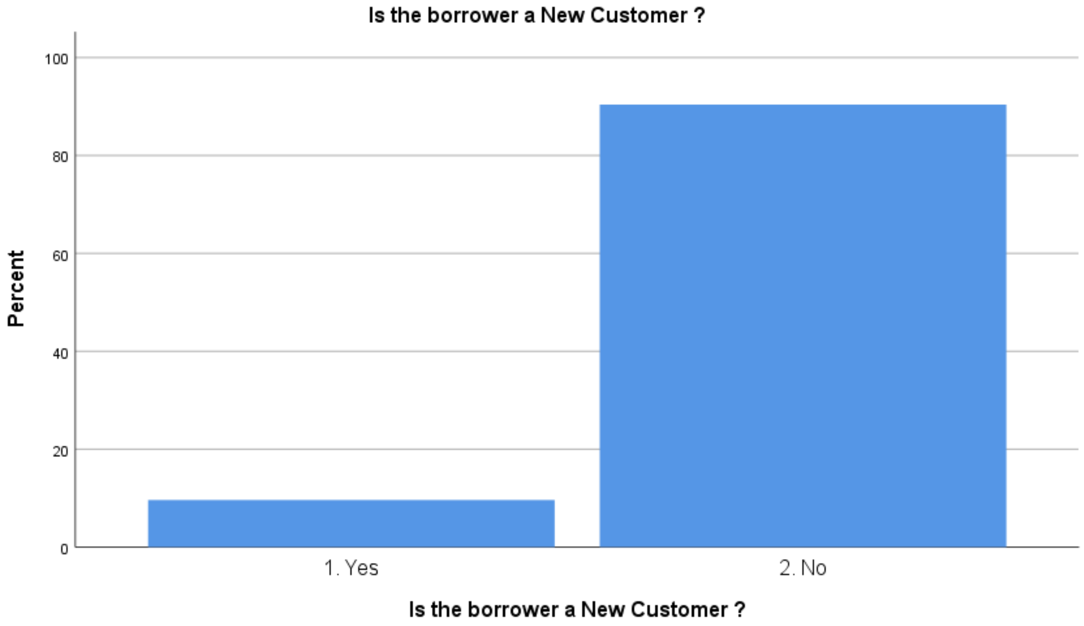


Figure 11: Is the borrower/ Applicant a new customer

Figure 11 shows that 90.4% of credit card applicants of bank of Ceylon has previous relationship with bank. There for can say the possibility of applying credit cards of new customers has only 9.6%.

When considering Is there any facility previously at any financial Institution figure 12 shows that 81.2% of applicants are previously held credit facilities in other financial institutions. Only 18.8% applicants are fresh applicants that not apply or held any credit facilities of other institutions.

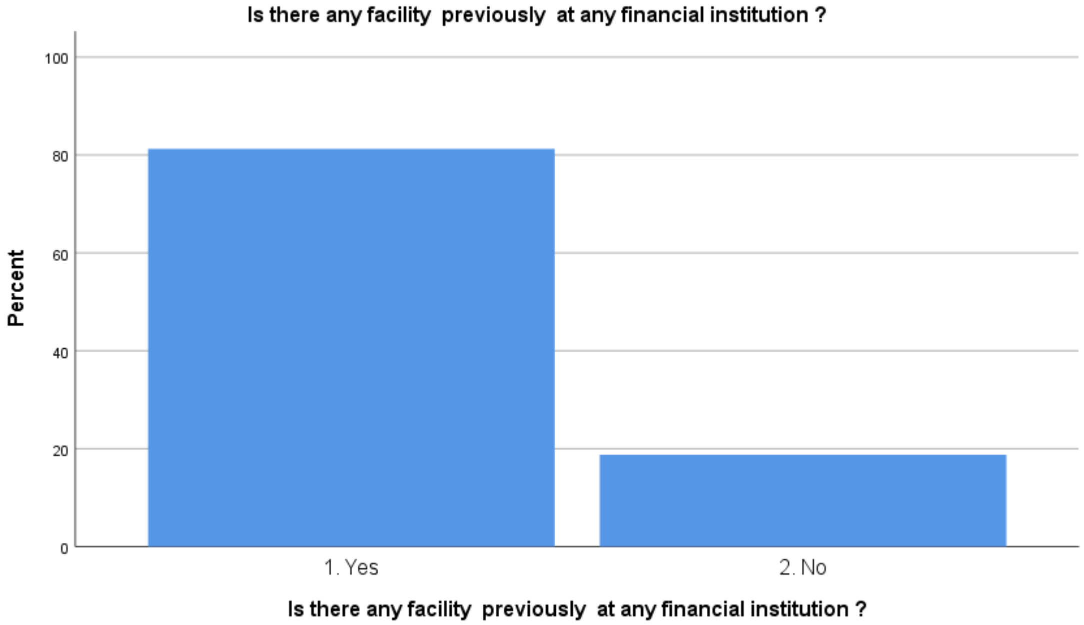


Figure 12: Is there any facility previously at any financial institution

When considering Has applicant ever guaranteed any facility at any financial Institution figure 13 shows that 73.9% of applicants has appear as guarantee in other financial institutions previously.

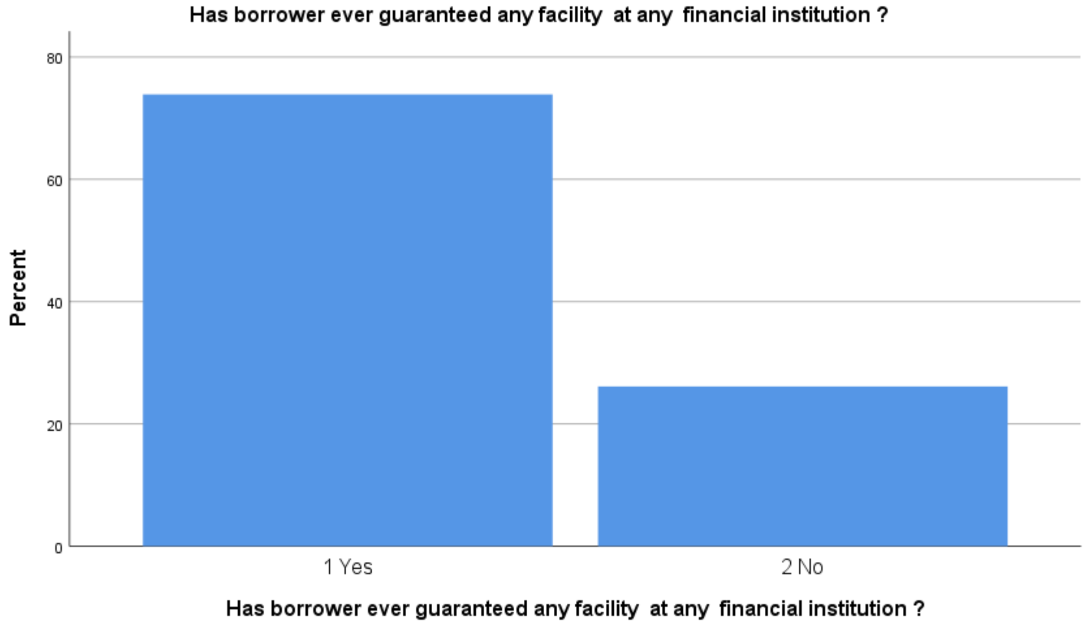


Figure 13: Has borrower ever guaranteed any facility at any institution

Considering about number of credit cards held figure 14 shows that 69.1% of applicants does not own any credit card. They are applying first time and 27.1% of applicants own cards less than or equal to 3 and 3.8% of applicants own more than 3 credit cards.

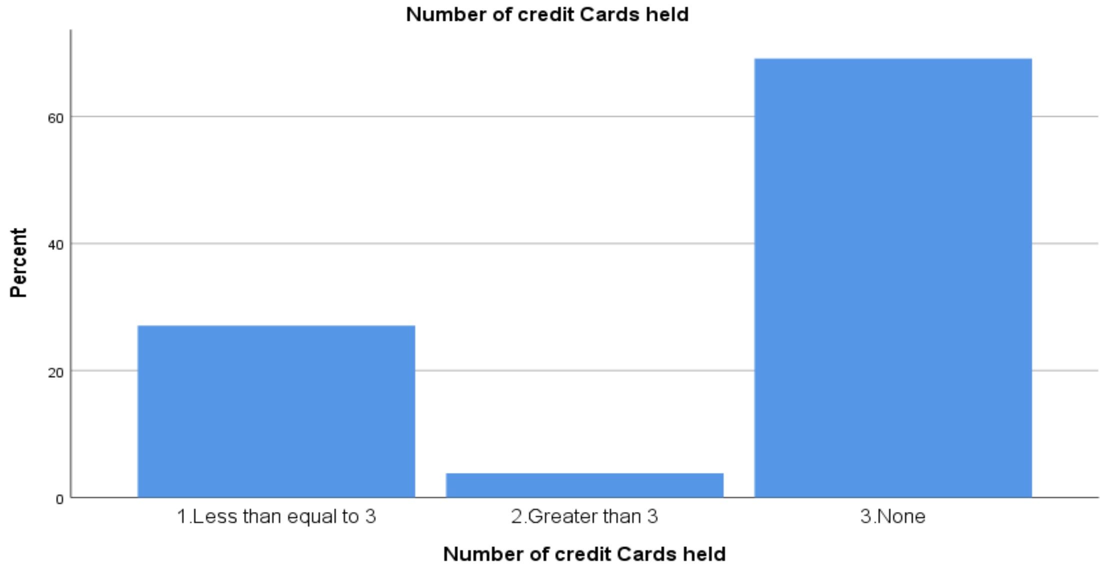


Figure 14: Number of Credit Cards Held

When considering Relationship with the Bank (Satisfactory running of SA/CA Accounts) figure 15 shows that 88.6% of applicants that operate account more than two years are applying for credit card. All other categories represent 11.4%. therefore, we can say applicants that maintain 24months willing to apply credit cards from Bank of Ceylon.

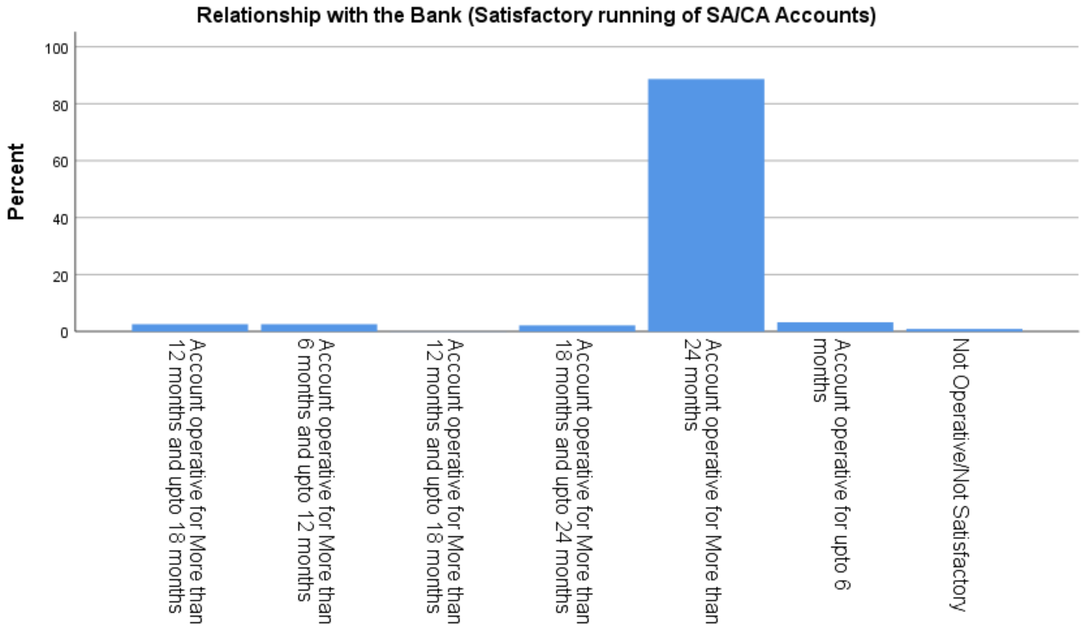


Figure 15 :Relationship with Bank (Satisfactory running of SA/CA Accounts)

When considering CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a borrower, CRIB - Conduct on Previous/Existing Loan facilities (if any) for the last 2 Years (including facilities at other financial institutions) - Appearing as a guarantor

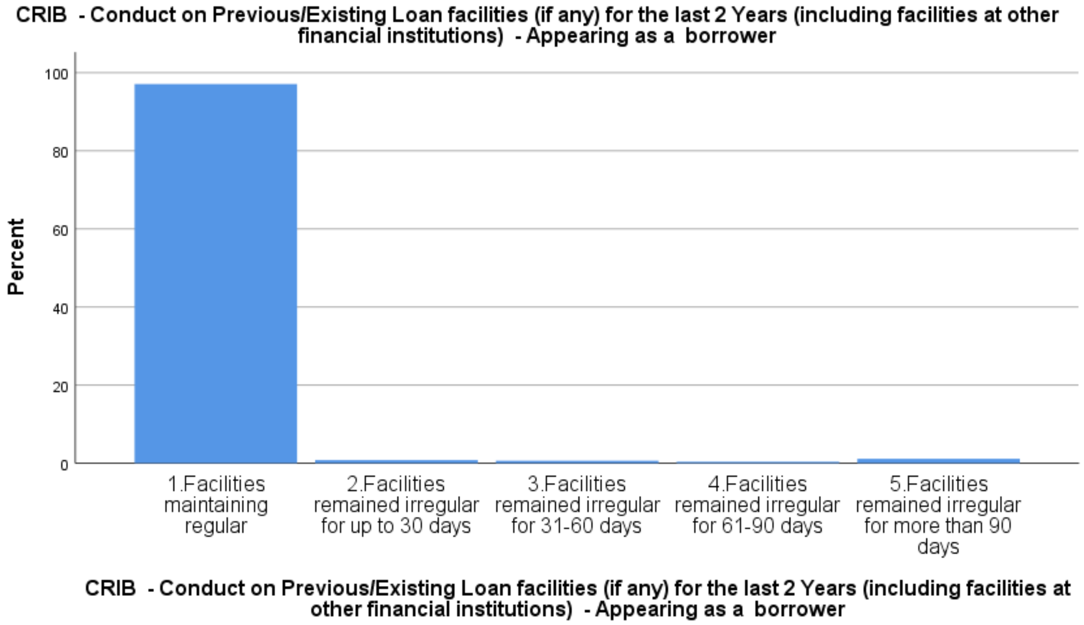


Figure 16: CRIB as Borrower

Figure 16 shows that 97.1% of applicants has good CRIB report as borrower. It means risk of default is less. According to figure 17 shows that 68.1% of applicants that signed as guarantors that maintain good CRIB report.

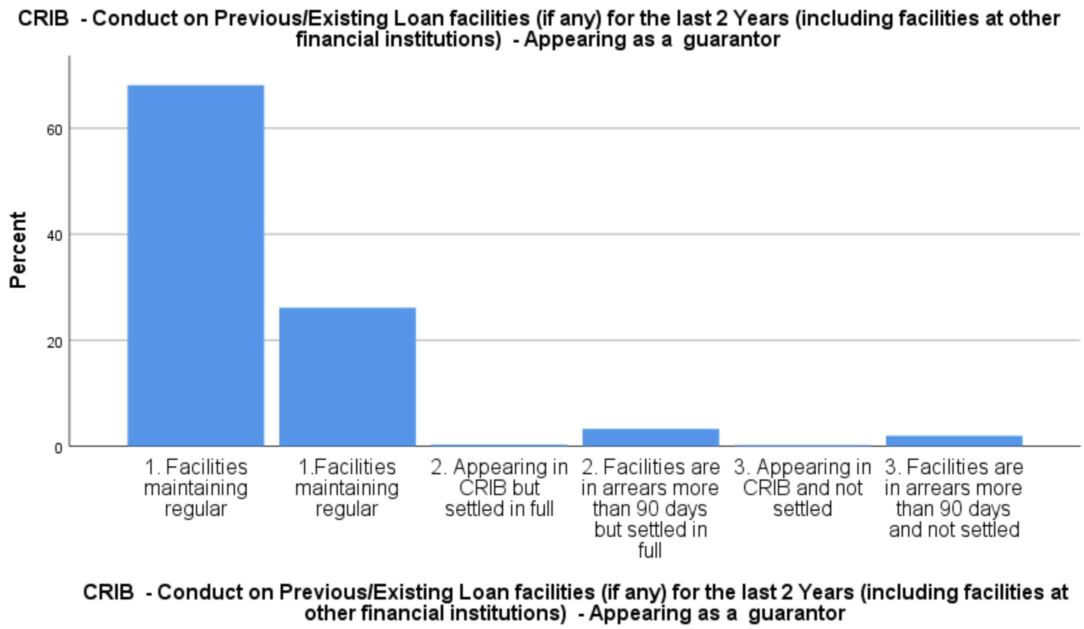
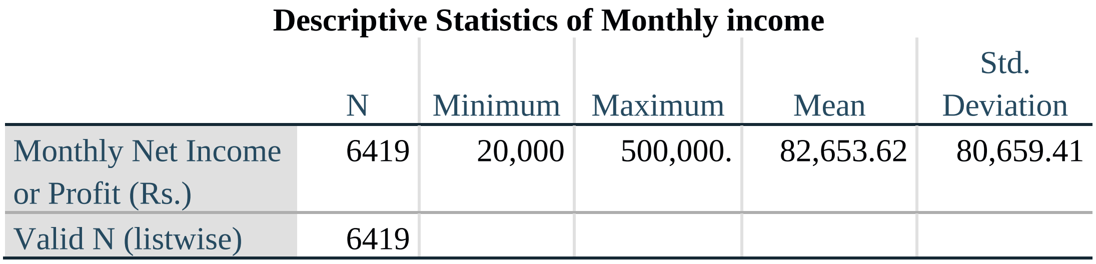


Figure 17: CRIB as Guarantor

Figure 18 shows the Monthly Net Income or Profit (Rs.) according to it minimum income level is 20,000 rupees and maximum income level is 500, 000 rupees mean income level is 82653.62 and standard deviation is 80659.41 rupees.



Descriptive Statistics of Monthly income

	N	Minimum	Maximum	Mean	Std. Deviation
Monthly Net Income or Profit (Rs.)	6419	20,000	500,000.	82,653.62	80,659.41
Valid N (listwise)	6419				

Figure 18: Monthly income or Profit (Rs.)

4.3.3 Analysis of Status of the Credit Card

In this project using machine learning algorithms analyze the relationship of possibility of categorize as Non-Performing Advances of credit cards in application of Bank of Ceylon. Normally in banking sector credit card Non performing level is between 5 % to 10% level in Sri Lankan credit card market. Figure 19 shows that NPA and Regular percentages of this data set.

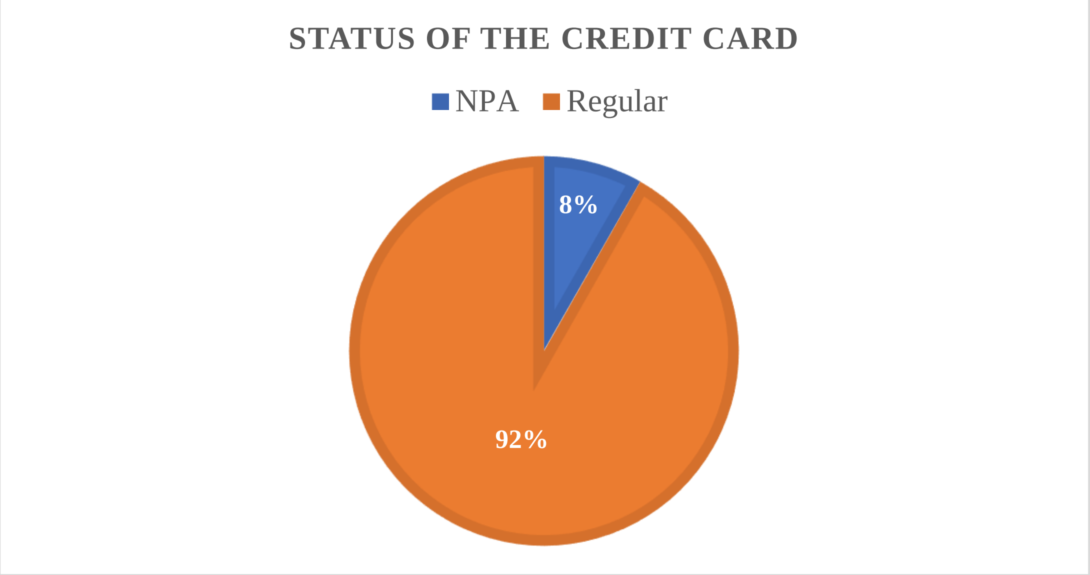


Figure 19: Status of the Credit Card

According to Figure 19 shows that 92% of cards in regular section and 8% of cards in NPA section. And this analysis shows that this data set is representing actual percentages of industry norms.

4.4 Result of the used prediction models

4.4.1 Used prediction models

In this study used supervised machine learning models to classify and predict applied Credit Card is going to non - performing in the future. For this analysis used K-Neighbors, Random Forest, Logistic Regression, Support Vector Machine and Naïve Bayes classifier. Using those algorithms train the classifier and build the classification model.

4.4.2 Accuracy of classifiers according to Confusion Matrix

Accuracy of the prediction classifiers calculated using confusion Matrix and Table 02 shows the results for each classifier.

Classifier	Accuracy
Logistic Regression Classifier	92.60%
K-Neighbors Classifier	91.27%
Random forest Classifier	92.52%
Naïve Bayes classifier	92.60%
Support Vector Machines classifier	**92.67%**

Table 2: Accuracy of Classifiers

According to Table 02 shows that support vector machines classifier has highest accuracy when comparing classifiers used in this study it has 92.67% of accuracy and logistic regression and Naïve Bayes classifiers has 92.60% accuracy according to confusion matrix. K- Neighbors classifier has the lowest accuracy of 91.27%.

All of classifiers shows the accuracy over 91%. Because this data set is biased data set. Banks issue credit cards for the customers that has ability for repay the dues. Otherwise, they do not issue cards. As a result of the NPA customer percentage less than the 8% of total issued cards. Therefore, this data set has biasness towards the Performing customer category. Therefore, we looking for correctness of the measurement of separability using AUC (Area under the curve analysis).

4.4.3 Area Under the Curve (AUC) Analysis

In preparation of machine learning model performance measurement is an essential task. In this study used the classification model to predict credit card defaults in future. For measurement of performance of this model used AUC- ROC curve (Area Under the Curve) (Receiver Operating Characteristics). This one is the most important evaluation metrics for checking any classification model's performance.

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

Following paragraphs analyzed about AUROC of each algorithm used in this model for calculate proper prediction of Credit Cards defaults.

4.4.3.1 AUC-ROC in relating to Random Prediction

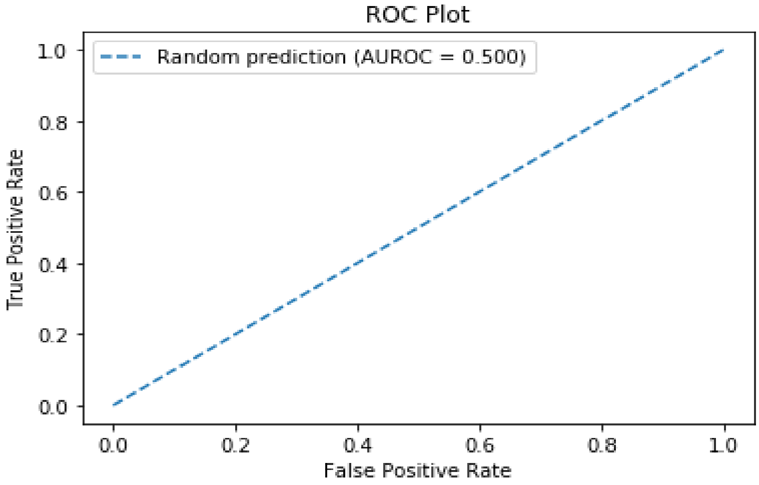


Figure 20: AUROC - Random Prediction

Figure 20 shows the AUROC in relating to random prediction. This curve used to compare the accuracy and reliability of the performance of the used algorithms. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly. If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives. When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives. When AUC=0.5, then the classifier is not able to distinguish between Positive and Negative class points. Meaning either the classifier is predicting random class or constant class for all the data points.

So, the higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.

4.4.3.2 AUC-ROC in relating to Logistic Regression classifier

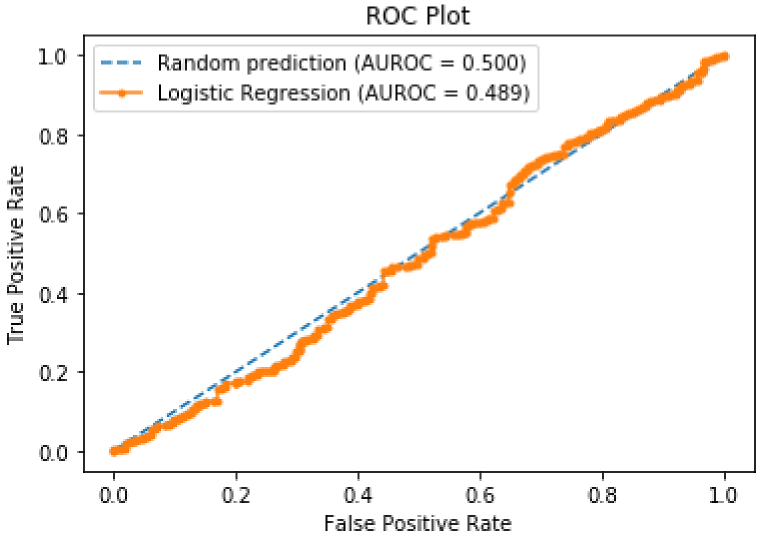


Figure 21: AUROC - Logistic Regression classifier

According to Figure 21 shows that AUC of the logistic regression classifier. AUC is less than the 0.500. When 0.5 < AUC < 1, there is a high chance that the classifier will be able to distinguish the positive class values from negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives. But in this Logistic regression classifier AUC is less than 0.500 it is 0.489 it means in this classification model that predicting the Credit card defaults Logistic regression classifier is not able to distinguish between positive and negative class points correctly. That means this classifier is not useful for this data set for predicting the Credit Card Defaults in future.

4.4.3.3 AUC-ROC in relating to Random Forest classifier

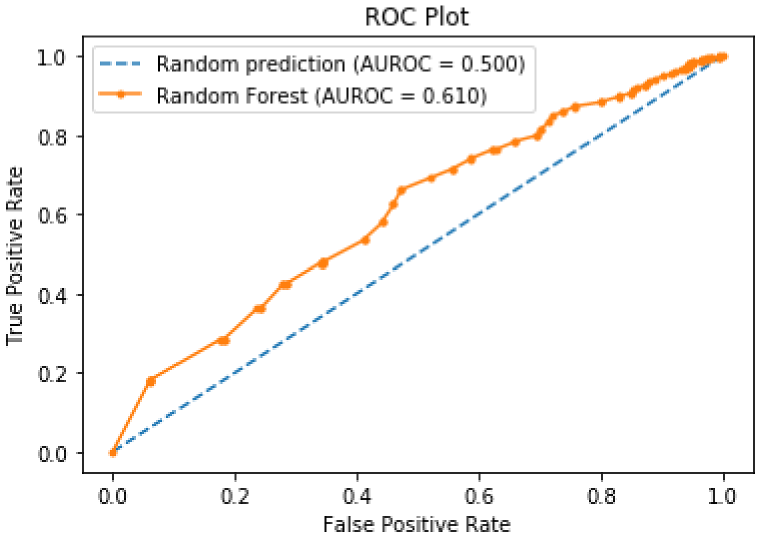


Figure 22: AUROC - Random Forest classifier

According to Figure 22 shows that AUC of the Random Forest Classifier. AUC is higher than the 0.500. When 0.5 < AUC < 1, there is a high chance that the classifier will be able to distinguish the positive class values from negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives. In this Random Forest classifier AUC is higher than 0.500 it is 0.610 it means in this classification model that predicting the Credit card defaults Random Forest classifier is able to distinguish between positive and negative class points correctly than the previous classifier Logistic regression. That means this classifier is useful for this data set for predicting the Credit Card Defaults in future.

4.4.3.4 AUC-ROC in relating to Naïve Bayes classifier

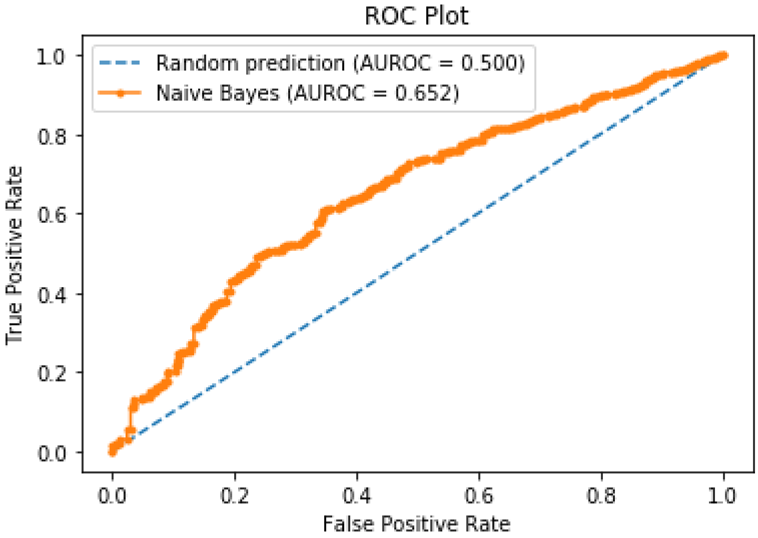


Figure 23: AUROC - Naive Bayes classifier

According to Figure 23 shows that AUC of the Naïve Bayes classifier. AUC is higher than the 0.500. When 0.5 < AUC < 1, there is a high chance that the classifier will be able to distinguish the positive class values from negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives. In this Random Forest classifier AUC is higher than 0.500 it is 0.652 it means in this classification model that predicting the Credit card defaults Naïve Bayes classifier is able to distinguish between positive and negative class points correctly than the previous classifiers Logistic regression, and Random Forest. That means this classifier is useful for this data set for predicting the Credit Card Defaults in future. This classifier has more accuracy and performance when comparing to the previous tested two classifiers Logistic regression and Random Forest Classifiers.

4.4.3.5 AUC-ROC in relating to KNeighbors classifier

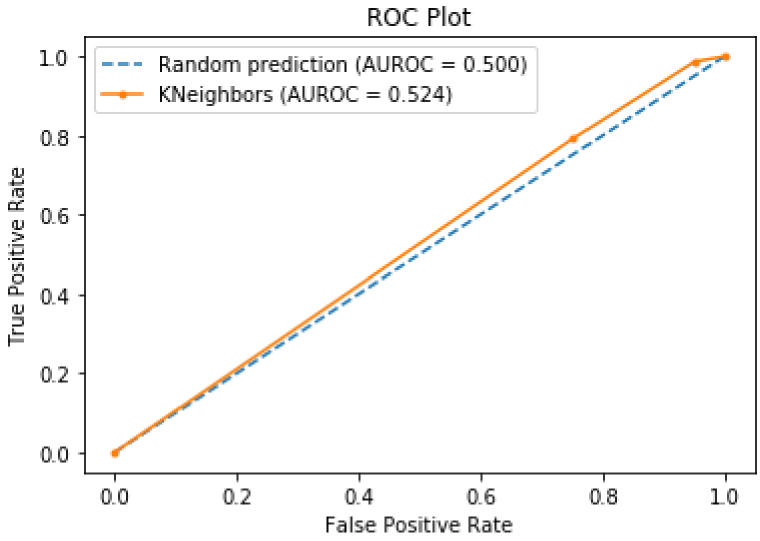


Figure 24:AUROC – KNeighbors Classifier

According to Figure 24 shows that AUC of the KNeighbors Classifier. AUC is higher than the 0.500. When 0.5 < AUC < 1, there is a high chance that the classifier will be able to distinguish the positive class values from negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives. In this KNeighbours classifier AUC is higher than 0.500 it is 0.524 it means in this classification model that predicting the Credit card defaults Random Forest classifier is not perfect to distinguish between positive and negative class points correctly (0.524 means same as 0.500). When AUC=0.5, then the classifier is not able to distinguish between Positive and Negative class points. Meaning either the classifier is predicting random class or constant class for all the data points.

4.4.3.6 AUC-ROC in relating to Support Vector Machines classifier

According to Figure 25 shows that AUC of the Support Vector Machines classifier. AUC is less than the 0.500. When 0.5 < AUC < 1, there is a high chance that the classifier will be able to distinguish the positive class values from negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives. But in this Support Vector Machines classifier AUC is less than 0.500 it is 0.489 it means in this classification model that predicting the Credit card defaults Support Vector Machines classifier is not able to distinguish between positive and negative class points correctly. That means this classifier is not useful for this data set for predicting the Credit Card Defaults in future.

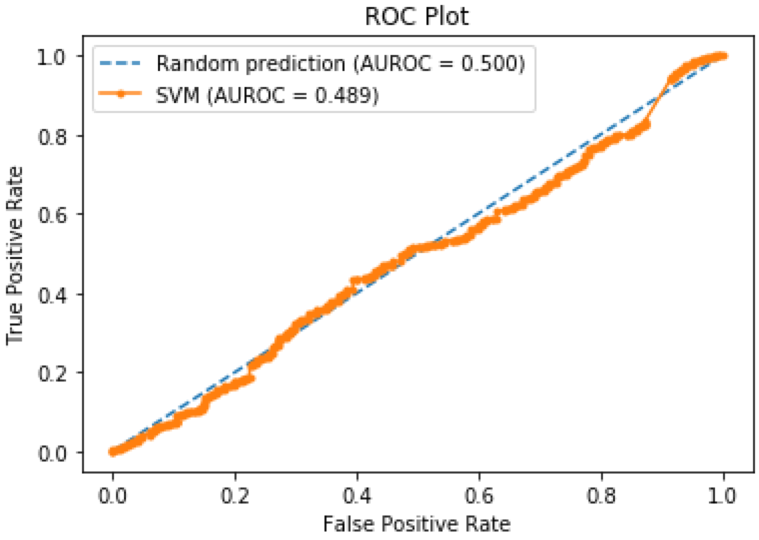


Figure 25: AUROC - Support Vector Machines Classifier

4.4.3.7 Comparison of AUC-ROC in relating to All Classifiers

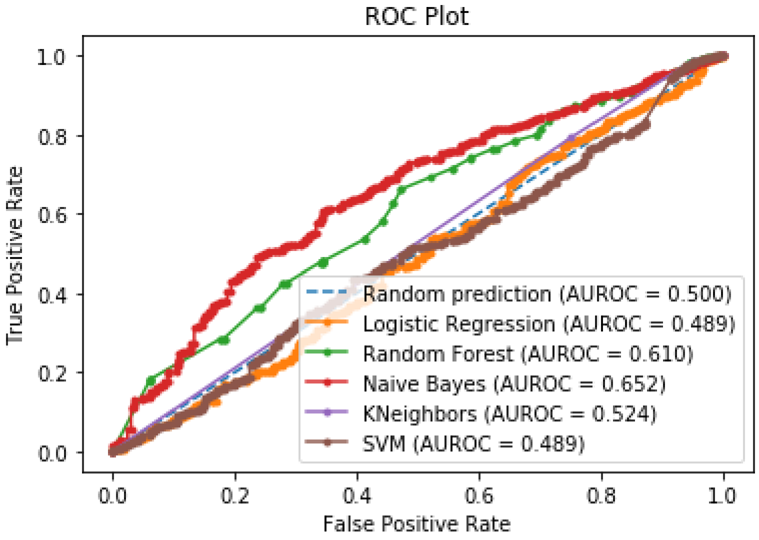


Figure 26: Comparison of AUC-ROC in relating to all classifiers

According to Figure 26 shows that AUC -ROC of the all classifiers used for this study and performance are showing in the Table 3.

Classifier	**AUC -ROC Value**
Logistic Regression	0.489
Random forest	0.610
Naïve Bayes	**0.652**
KNeighbors	0.524
Support Vector Machines	0.489

Table 3: Comparison of AUC-ROC Values

It is evident that plot that the AUC for the Logistic regression, Random Forest, Naïve Bayes, KNeighbors and support Vector Machines AUC for the Naïve Bayes ROC curve is higher than for the all-other curves comprising the value of 0.652. Therefore, according to above output can say that Naïve bayes classifier did a better job of classifying the positive class in the Credit card database of Bank of Ceylon. It means we can use Naïve Bayes classifier for predicting Credit Card defaults in the stage of application of the credit card accurately.

Chapter Five

Conclusion

5.1 Introduction and Conclusion

This research exposed that exposed Predicting Probability of Credit Card Default at the Stage of Credit Card Application Using Supervised Machine Learning Approaches. In this chapter discussed about what are the contributions to the knowledge, suggestions for future research and personal reflections of probability of Credit card default at the stage of credit card application using supervised Machine learning approaches

The use of machine learning techniques for the prediction of credit card defaulters is essential for the identification of credit risk. This can help the financial institutions in designing their future strategies. The proposed system uses Naive Bayes, logistic regression, SVM, KNeighbors and Random Forest on credit scoring data set in relating to Credit card applications of Bank of Ceylon. The performance of these classifiers is evaluated using the accuracy of its prediction. The classifier with the highest accuracy/ performance is found to be Naïve Bayes. It is followed by Random Forest and KNeighbours. Support vector machines and Logistic Regression gives less accuracy as compared to the other classifiers. The performance of classifiers is slightly reduced when large number of instances are given for classification.

According to this study we can conclude that for this data set relating to Credit Card of Bank of Ceylon can use Naïve Bayes Classifier for Better classification than the other classifications used called Logistic regression, KNeighbours, Support Vector Machines and Random Forest.

The findings and this model have emphasized that the proper prediction of non-performing loans and advances that can be arising in the future by using modern technologies (machine learning approaches) and then bankers can understand customer properly and provide accurate decision whether it is approving or not the applied credit card easily.

5.2 Contribution to Knowledge

This project can help banks in predicting the future of Credit Card non performing or not and its status and depends on that they can take action in initial stage of approving credit card. Using this application banks can reduce the number of non-Performing loans and from incurring sever losses. Several machine learning algorithms used for supervised learning called KNN, Random Forest, Logistic Regression, Support Vector Machines and Naïve Bayes were used to prepare the data and to build the classification model. Python Package libraries help in successful data analysis and feature selection. Using this methodology bank can easily identify the required information from huge amount of data sets and helps in successful Credit card prediction to reduce the number of Non performing percentage of Credit Cards. Data Mining and machine learning techniques are very useful to the banking sector for better targeting and acquiring new customers, most valuable customer retention, automatic credit card approval which is used for fraud prevention, fraud detection in real time, providing segment-based products, analysis of the customers, transaction patterns over time for better retention and relationship, risk management and marketing.

5.3 Suggestions and future researches

From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems. There have been numbers cases of computer glitches, errors in content and most important weight of features are fixed in automated prediction system, so in the near future the so – called software could be made more secure, reliable and dynamic weight adjustment. In near future this module of prediction can be integrate with the module of automated processing system. the system is trained on old training dataset in future software can be made such that new testing date should also take part in training data after some fix time.

In addition to that in this study used only supervised learning machine learning approaches to predict credit card defaults but researchers can use other machine learning approaches such as neural networks to predict credit card defaults in future.

Specially this study can directly plug into the loans and advances database of the bank and can predict probability of transfer loans and advances to non - performing sector of the bank. As a result of that banks can reduce their provisions requirement drastically and increase the revenue as well as profits by decreasing non - performing loans and advances portfolio and increasing the asset portfolio of the banks' balance sheet.

In addition to suggest to improve and fine tune this machine learning model using hyper parameters techniques select the best possible options and introduce new features to the bank credit scoring model. then accuracy of this model increased. For each classifier can fine tune each parameter using hyper parameters method.

5.4 Personal Reflection about study

In a new or emerging market, the operational, technical, business and cultural issues should be considered with the implementation of the credit scoring models for credit cards and retail loan products. The operational issues relate to the use of the model and it is imperative that the staff and the management of the bank understand the purpose of the model.

Application scoring models should be used for making credit decisions on new applications and behavioral models for credit cards and retail loan products to supervise existing borrowers for limiting the expansion or for marketing new products. The technical issues relate to the development of proper infrastructure, maintenance of historical data and software needed to build a credit scoring model for credit cards and retail loan products within the bank. The business issues relate to whether the soundness and safety of banks could be achieved through the adoption of quantitative credit decision models, which would send a positive impact in the banking sector. The cultural and demographic issues relate to making credit irrespective of race, colour, sex, religion, marital status, age or ethnic origin as well as other financial variables such as income, CRIB status, number of cards used and relationships with bank. Further, models have to be validated so as to ensure that the model performance is compatible in meeting the business as well as regulatory requirements. Thus, the above issues have to be considered while developing and implementing credit scoring models for credit cards and retail loan products.

References

A. Bellotti, J. Crook, 2009. Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Reserch Society,* Volume 60, pp. 1699-1707.

A. Goyal, R. Kaur, 2016. Accuracy Prediction for Loan Risk using Machine Learning Models. *International Journal of Computer Science Trends and Technology,* 4(1), pp. 52-57.

B. Baesens, D. Roesch, H. Scheule, 2016. *Credit Risk Analytics : Measurement Techniques, Applications, and Examples in SAS.* s.l.:Wiley .

B. Gultekin and E.B. Sakar, 2018. *Variable Importance Analysis in Default Prediction using Machine Learning Techniques.* Portugal, s.n., pp. 56-62.

C.J Nali´ and A. Švraka, 2018. *Using Data Mining Approaches to Build Credit Scoring Model.* Jahorina, s.n.

D.J. Hand, S.D. Jacka, 1998. *Statistics in Finance.* London: Hodder Education.

E. Agbemava, I.K. Nyarko, T.C. Adade, A.K. Bediako, 2016. Logistic Regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana. *European Scientific Journal,* 12(1), pp. 175-189.

F. Butaru, Q.Chen, B. Clark, S. Das, W. Andrew, 2016. Risk and Risk Management in the Credit Card Industry. *Journal of Banking and Finance,* Volume 72, pp. 218-239.

H. Abdou and J. Pointon, 2011. Credit scoring, statistical techniques and evaluation criteria: a review of the litreature. *Intelligent Systems in Accounting, Finance & Management,* 18(2-3), pp. 59-88.

H.A Abdou and J. Pointon, 2011. Credit scoring, statistical techniques and evaluation criteria: A review of the litreture. *Intelligent Systems in Accounting, Finance and Management,* 18(2-3), pp. 59-88.

H.C. Koh, W.C. Tan, C.P. Goh, 2006. A Two-step method to Construct Credit Scoring Models with Data Mining Techniques. *International Journal of Business and Information,* 1(1), pp. 96-118.

J. Tejaswini, M.T. Kavya, N.D.R. Ramya, S.P. Triveni, R.V. Maddumala, 2020. Accurate Loan Approval Prediction based on Machine Learning Approach. *Journal of Engineering Science*, 11(4), pp. 523-532.

J.A Hamid and M.T. Ahmed, 2016. Developing Prediction Model of Loan Risk in Banks using Data Mining. *International Journal on Machine Learning and Applications*, 3(1), pp. 1-9.

L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, 1984. *Classification and Regression Trees.* New York: Chapman & Hall/CRC.

L.A. Tudor, A. Bara, V.S. Opera, 2017. Comparative analysis of data mining methods for predicting credit default probabilities in a retail bank portfolio. *Latest Trends in Information Technology*, pp. 117-122.

M.Jayadev, N.M. Shah, R. Vadlamani, 2019. *Predicting Educational Loan Defaults: Application of Artificial Intelligence Models.* [Online] Available at: https://www.iimb.ac.in/sites/default/files/2019-12/WP%20No.%20601.pdf
[Accessed 18 October 2020].

M.M. Hassan and T. Mirza, 2020. Credit Card Default Prediction Using Artificial Neural Networks. *GIC Science Journal*, 7(7), pp. 383-390.

N. Madane and S. Nanda, 2019. Loan Prediction analysis using decision Tree. *Journal of the Gujarat Research Society*, 21(14s), pp. 214-221.

N. Setiawan, Suharjito, Diana, 2019. A Comparison of Prediction Methods for Credit Default on Peer to Peer Lending using Machine Lerning. *Procedia Computer Science,* Volume 157, pp. 38-45.

N. Torvekar and S.P. Game, 2019. Predictive Analysis of Credit Score for Credit Card Defaulters. *International Journal of Recent Technology and Engineering,* 7(5S2), pp. 283-286.

N.G. Matthew and S.S Boateng, 2013. Credit risk and loan default among Ghanaian banks: An exploratory study. *Management Science letters,* 3(3), pp. 753-762.

O.M. Faruk, S.M. Islam, n.d. *An Analytical Review of Non Performing Loan Bangladesh and Global Perspectives.* [Online] Available at: https://www.academia.edu/31657465/An_Analytical_Review_of_Non_Performing_Loan_Bangladesh_and_Global_Perspectives
[Accessed 15 October 2020].

P. Supriya, M. Pavani, N. Saisushma, V.N. Kumari, K. Vikas, 2019. Loan Prediction by using Machine Learning Models. *International Journal of Engineering and Techniques,* 5(2), pp. 114-148.

R. Azam, M. Danish, S. Akbar, 2012. *The significance of socioeconomic factors on personal loan decision a study of consumer banking local private banks in Pakistan.* [Online]
Available at: https://mpra.ub.uni-muenchen.de/42322/
[Accessed 15 October 2020].

R. Mbuvha, I. Boulkaibet, T. Marwala, 2019. *Automatic Relevance Determination Bayesian Neural Networks for Credit Card Default Modelling.* [Online] Available at: https://www.semanticscholar.org/paper/Automatic-Relevance-Determination-Bayesian-Neural-Mbuvha-Boulkaibet/750ede618de87a952e3842e804cd96e4e18ca9ad#citing-papers
[Accessed 20 October 2020].

R. Patibandla et al, 2017. Significance of Embedded Systems to IoT. *International Journal of Advance Engineering and Research Development,* 16(2), pp. 860-867.

R.A. Itoo, A. Selvarasu, A.J. Filipe, 2015. Loan Product and Credit Scoring by commercial Banks (India). *International Journal of Latest Trends in Finance & Economic Sciences,* 5(1), pp. 851-860.

S. Neema and B. Soibam, 2017. The comparison of machine learning methods to achieve most cost-effective prediction for credit card default. *Journal of Management Science and Business Intelligence,* 2(2), pp. 36-41.

S.R. Islam, W. Eberle, S.K. Ghafoor, 2019. *Credit Default Mining Using Combined Machine Learning and Heuristic Approach.* [Online]

Available at: https://arxiv.org/ftp/arxiv/papers/1807/1807.01176.pdf [Accessed 18 October 2020].

T. Chou and M. Lo, 2018. Predicting Credit Card Defaults with Deep Lerning and other Machine Lerning Models. *International Journal of computer Theory and Engineering,* 10(4), pp. 105-110.

T.S. Lee, C.C. Chiu, Y.C. Chou, C.J. Lu, 2006. Mining the customer credit rating using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics and Data Analysis,* Volume 50, pp. 1113-1130.

U. Aslam, T.I.H. Aziz, A. Sohail, K.N. Batcha, 2019. An empirical study on loan default prediction models. *Journal of Computational and Theoritical Nanoscience,* 16(8), pp. 3483 - 3488.

X.L. Li, Y. Zhong, 2012. An overview of Personal Credit Scoring : Techniques and Future Work. *International Journal of Intelligence Science,* Volume 2, pp. 181-189.

Y. Hou and D. Dickinson, 2008. *The Non-performning Loans : Some Bank - Level Evidences.* France, s.n.

Yang, Y., 2007. Adaptive Credit Scoring with kernal learning methods. *European Journal of Operational Research,* 183(3), pp. 1521-1536.

Annexure

Annexure 1: Codes

Import Libraries and Split Data Set

```
import pandas as pd
import numpy as np
import pickle
cc = pd.read_csv('data2.csv')
# Import train_test_split
from sklearn.model_selection import train_test_split

# Segregate features and labels into separate variables
X,y =  cc.loc[:, cc.columns != 'STATUS'], cc.STATUS

# Split into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X,
                                   y,
                                   test_size=0.20,
                                   random_state=42)

from sklearn.linear_model import LogisticRegression
# Instantiate a LogisticRegression classifier with default parameter values
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
# Import confusion_matrix
from sklearn.metrics import confusion_matrix

# Use logreg to predict instances from the test set and store it
y_pred = logreg.predict(X_test)
```

```
# Get the accuracy score of logreg model and print it
print("Accuracy of logistic regression classifier: ", logreg.score(X_test, y_test))

# Print the confusion matrix of the logreg model
confusion_matrix(y_pred, y_test)
logreg.score(X_test, y_test)
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn import svm
#rf = RandomForestClassifier(max_features=5, n_estimators=500)
rf = RandomForestClassifier(n_estimators=100)
rf.fit(X_train, y_train)
nb = GaussianNB()
nb.fit(X_train, y_train)
nb.score(X_test, y_test)
kn = KNeighborsClassifier(n_neighbors=3)
kn.fit(X_train, y_train)
kn.score(X_test, y_test)
sv = svm.SVC(probability=True)
sv.fit(X_train, y_train)
sv.score(X_test, y_test)
### Prediction probabilities

r_probs = [0 for _ in range(len(y_test))]
lr_probs = logreg.predict_proba(X_test)
rf_probs = rf.predict_proba(X_test)
nb_probs = nb.predict_proba(X_test)
kn_probs = kn.predict_proba(X_test)
```

```
sv_probs = sv.predict_proba(X_test)
lr_probs = lr_probs[:, 1]
rf_probs = rf_probs[:, 1]
nb_probs = nb_probs[:, 1]
kn_probs = kn_probs[:, 1]
sv_probs = sv_probs[:, 1]
from sklearn.metrics import roc_curve, roc_auc_score
r_auc = roc_auc_score(y_test, r_probs)
lr_auc = roc_auc_score(y_test, lr_probs)
rf_auc = roc_auc_score(y_test, rf_probs)
nb_auc = roc_auc_score(y_test, nb_probs)
kn_auc = roc_auc_score(y_test, kn_probs)
sv_auc = roc_auc_score(y_test, sv_probs)
print('Random (chance) Prediction: AUROC = %.3f' % (r_auc))
print('logistic regression: AUROC = %.3f' % (lr_auc))
print('Random forest: AUROC = %.3f' % (rf_auc))
print('Naive Bayes: AUROC = %.3f' % (nb_auc))
print('KNeighbors : AUROC = %.3f' % (kn_auc))
print('SVM : AUROC = %.3f' % (sv_auc))
r_fpr, r_tpr, _ = roc_curve(y_test, r_probs)
lr_fpr, lr_tpr, _ = roc_curve(y_test, lr_probs)
rf_fpr, rf_tpr, _ = roc_curve(y_test, rf_probs)
nb_fpr, nb_tpr, _ = roc_curve(y_test, nb_probs)
kn_fpr, kn_tpr, _ = roc_curve(y_test, kn_probs)
sv_fpr, sv_tpr, _ = roc_curve(y_test, sv_probs)
import matplotlib.pyplot as plt
plt.plot(r_fpr, r_tpr, linestyle='--', label='Random prediction (AUROC = %0.3f)' %
r_auc)
plt.plot(lr_fpr, lr_tpr, marker='.', label='Logistic Regression (AUROC = %0.3f)' %
lr_auc)
plt.plot(rf_fpr, rf_tpr, marker='.', label='Random Forest (AUROC = %0.3f)' % rf_auc)
plt.plot(nb_fpr, nb_tpr, marker='.', label='Naive Bayes (AUROC = %0.3f)' % nb_auc)
plt.plot(kn_fpr, kn_tpr, marker='.', label='KNeighbors (AUROC = %0.3f)' % kn_auc)
plt.plot(sv_fpr, sv_tpr, marker='.', label='SVM (AUROC = %0.3f)' % sv_auc)

# Title
plt.title('ROC Plot')
# Axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# Show legend
plt.legend() #
# Show plot
plt.show()
```