Credit Card Approval Prediction by Using Machine Learning Techniques

M. P. C. Peiris

2019

Credit Card Approval Prediction by Using Machine Learning Techniques

A dissertation submitted for the Degree of Master of Business Analytics

M. P. C. Peiris

University of Colombo School of Computing

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: Makumburage Poornima Chathurangi Peiris

Registration Number: 2018/BA/026

Index Number: 18880269

PCPeiris.

Signature: Date: 14/09/2021

This is to certify that this thesis is based on the work of Ms. Poornima Peiris under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. Rushan Abeygunawardana

Signature: Date: 14/09/2021

I would like to dedicate this thesis to:

My Mother

My Grandmother

My Uncle

My Relatives

My Teachers at UCSC

My Friends

Freshers in Data Science arena around the world

ACKNOWLEDGEMENTS

It is with pleasure that I express my heartfelt gratitude to Dr. Rushan Abeygunawardana, my faculty supervisor who helped and guided me throughout this endeavor.

I would like to thank My mother Iranganie, uncle Lakshman, grandmother and best friends for their love, support and understanding during the years of my education.

Lastly, I offer my gratitude to all of those who assisted me in providing information and extending their support toward completing this research project successfully.

ABSTRACT

This research is focusing on application of machine learning (ML) techniques to predict customer eligibility for a credit card to mitigate possible future credit risk which may affect the bank's financial stability and credit performance. Credit card is a credit facility given for a customer by banks and finance companies around the globe. The credit facility has a credit risk for the banks and financial companies. The repayments are least assured and it often ends up as a non-performing credit facility (NPL). To mitigate credit risk banks are assessing applicant's creditworthiness and checking the eligibility before granting a credit facility. The decision is mostly based on traditional credit scoring models and credit worthiness will not always be accurate. This project aims to help banking and financial institutions to identify and interact with creditworthy customers by using predictive models. We used Artificial Neural Network (ANN) and Support Vector Mechanism (SVM) to develop models. Under ANN we have tested models using different sizes of batches, low and high learning rates. Linear SVM and Nonlinear SVM both models used to evaluate the best SVM method. Statistical methods under filter-based feature selection methods applied for feature selection. Model accuracy checked using Mean Absolute Error, Confusion Matrix, Area Under Curve (AUC) for training and test data. We have evaluated three classifiers and we observed that Nonlinear SVM is performed better than ANN and linear SVM. Nonlinear SVM model Accuracy is 0.88, Precision is 0.88, Recall is 0.90 and AUC is 0.89. Accuracy, Precision and Recall values are higher in Nonlinear SVM than ANN and Linear SVM. Recall rate is 0.90 means the model predicts positive class 90% correctly. We also realized that customer behavior might be different from country to country and application of several real banking datasets not limited to customer demographic and socio-cultural but also other credit facility features including COVID-19 impact to be an area of concern for researchers. Furthermore, whether there is a relationship between Nonlinearity in highly imbalanced class problems with SMORTE application is another area of concern for researchers

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Project Overview

This research is focusing on application of machine learning (ML) techniques to predict customer eligibility for a credit card.

One of key objective of the bank is to increase the returns. When increasing the returns there is an increase of risk. Banks are faced with various risks such as interest rate risk, market risk, credit risk, off-balance-sheet risk, technology and operational risk, foreign exchange risk, country or sovereign risk, liquidity risk, liquidity risk and insolvency risk. Effective management of these risks is key to a bank's performance. Credit can be defined as the risk of potential loss to the bank if a borrower fails to meet its obligations (interest, principal amounts). Continuously monitoring of customer payments could reduce the probability of accumulating non-performing assets (NPA). Whether to grant or not to grant a loan to a customer is one of key decisions of banks use to reduce probable NPA at the first hand. Credit card as a credit facility instruments banks need to effectively managed credit risk of the facility. The Basel Accord allows banks to take the internal ratings-based approach for credit risk. Banks can internally develop their own credit risk models for calculating expected loss.

There are several manual steps involving when granting a credit card to a customer. Assessing applicant's creditworthiness and checking the eligibility are the key factors and decisions the bank would take about a credit worthiness will not always be accurate. Application of machine learning techniques can eliminate manual paperwork, time-consuming processes and most importantly data driven decision making before granting a credit card to a customer. In this research, different supervised machine learning algorithms were used to develop models and follow the steps in cross-industry standard process for data mining (CRISP-DM) life cycle. Accuracy of models was validated by using different validation techniques.

1.2 Motivation

In times of yore, when providing a credit card to a customer, banks had to rely on the applicant's background and the history to understand the creditworthiness of the applicant. The process includes scrutinisation of application data with reference documents and this process was not always accurate and customers and the bank had to face difficulties in approving the credit card. But with the digital transformation, there is a growth in Artificial Intelligence & Machine

Learning Technology in the past two decades. Therefore, ML techniques being used to evaluate credit risk and automate credit scoring by predicting the customer eligibility correctly using customer demographic data and historical transactional data. Furthermore, ML helps banks to make smarter data –driven decisions for customers; use banking data in a more productive and efficient way; streamline customer interaction by removing manual and lengthy processes.

1.3 Statement of the problem

Many researchers have conducted machine learning applications on credit scoring and customer default predictions. Researchers' have concluded that SVM (support vector machine) and ANN (Artificial Neural Network) performed better than other classifiers. However, it is important to study how these two algorithms behave differently with filter based feature selection and balancing imbalanced data which is inherited by nature using Synthetic Minority Oversampling Technique (SMORTE).

"To examine two algorithms and identify best classification algorithm to predict customer eligibility for a credit card and to minimize possible credit loss "

1.4 Research Aims and Objectives

The primary focus of the research is expressed under aims and objectives as follows.

1.4.1 Aim

This research supports the decision making process while speeding up the process to give a benefit for the bank as well as for the applicant and to attract on time paying customers by using banking data for smarter data–driven decision making. This research is highly applicable for Sri Lankan banking industries as most of the banks are granting credit card facilities to the customers. Hence the application of the model to local context to be considered.

1.4.2 Objectives

Research objectives of the project as follows:

- To predict the customer eligibility for a credit card to minimize possible future credit loss by using supervised machine learning techniques.

1.5 Background of the Study

Commercial banks contribute to economic growth in various aspects. One of the biggest revenue streams of any banking or financial institution would be from the interest charged from the lending. Banks have to face the biggest credit risk in all their lending. There are various lending products the banks are offering to the customers. However, Credit cards are one of the key lending products any bank would ever have. Almost all the financial institutions across the globe are going through challenging time and credit risk in offering credit facilities to their end customers. The repayments are least assured and it often ends up as a non-performing credit facility (NPL). This will in return affect banks cash flow and leads to build up backlogs in balance sheet which will not look good if the bank is a listed organization.

Banks and financial institutions are critically assessing eligibility for a credit facility before granting facility to the customer due to the credit risk factor the credit card involved in. This process involves verification, validation, and approval and may cause delay of granting a facility which will be disadvantageous for the applicant as well as for the bank. Credit officers determine whether the borrowers can fulfill their requirements to being eligible for a facility and these judgments and predictions are always not accurate. Credit scoring is a traditional method assessing the credibility of a customer / entity applying for a bank credit facility. How much ever the banks and financial institutions are doing the background check of the individual customers by analyzing their eligibility, the bank most of the time end up in making wrong decisions. The study determines whether an Artificial Intelligence system using Machine Learning Technology can assist the industry in overcoming from this risk.

1.5.1 What Is a Credit Card?

Credit card is a credit facility given for a customer by banks and finance companies. It has a higher annual percentage rate (APR) than other consumer loans. By law, card issuers must provide 21 days of grace period before interest on purchases and begin to accrue. When customers paying off balance before the grace period expired consider as a good practice. Interest charges will begin for any unpaid balance typically after one month of purchase is made. In case of any unpaid balance left it had been carried forward from a previous month and for new charges there is no grace period provided. Interest will be accruing daily or monthly according to issuer interest and the country's financial policies (Thomas J. Catalano, 2020).

Credit card will be entered to delinquent state if the customer failed to paid minimum monthly amount for 30 days from original due date. Most of financial institutes start to reaching customers when customer card status become past due. After 60 days or more delinquent status become overdue and most companies involve in taking legal actions to start debt collection (Fernando, 2021).

1.5.2 Component of a Credit Card

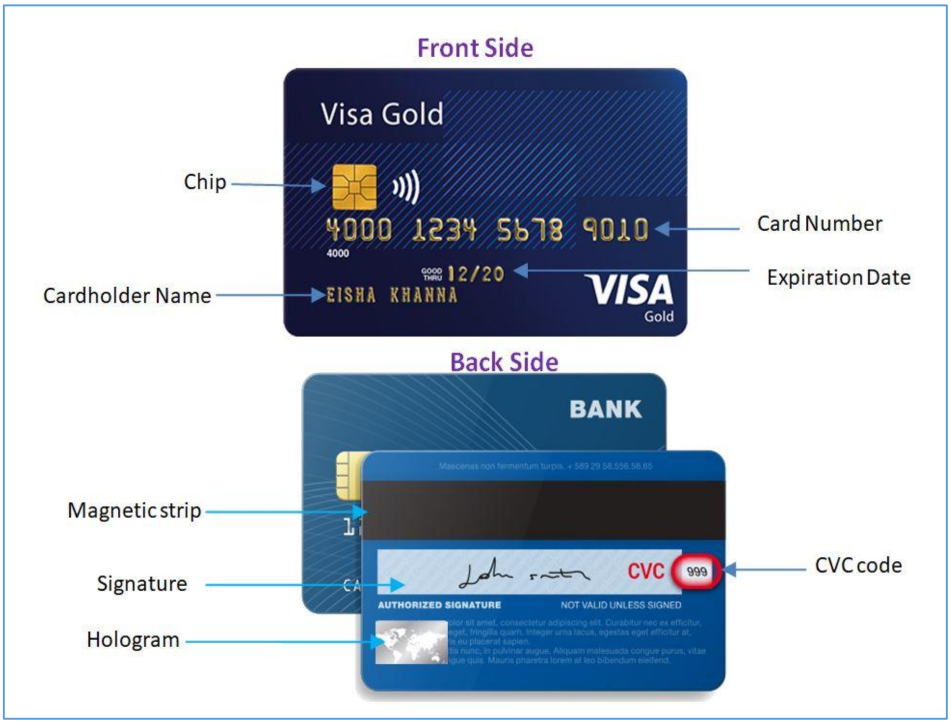


Figure 1.1 - Component of a Credit Card

Figure 1.1 illustrated components of a credit card and details of components were listed below.

- **Issuer Logo:** In front of the credit card, credit card network logo (e.g. visa, master) and issuing bank logo displayed.
- **EMV Chip:** The chip stores card data in an encrypted way to prevent stealing of credit card number easily.
- **Magnetic Strip:** The magnetic strips are readable through some specific machines used for monetary transactions. Also it contains account data.

- **Card Holder Name and Card Number:** Card holder name & Credit Card number appeared in front side of the card.
- **Credit Card Expiration Date**: Card has an expiration date. The date shows the month and the year and helps merchants to identify the validity of the card.
- **Signature Box:** Signature box is the place cardholders are supposed to place their signature.
- **CVV Code:** In back side of the card there is CVV number. It is three-digit combination and used to protect customers' financial transaction from fraud and theft.
- **Hologram:** In the backside of card unique three-dimensional hologram display of credit card network. (E.g. Visa uses a dove hologram, MasterCard – a world map)

1.5.3 Credit Line

A line of credit (LOC) is a stipulated amount of money that a card issuer has agreed to lend for a customer at the beginning of credit card account opening. Until the limit is reached, the borrower can draw money from the credit card and as money is repaid, it can be borrowed again in the case of an open line of credit. Credit line can be increase after evaluating customers' repayment capacity later.

1.5.4 Types of Credit Cards

Most popular credit card networks/brands are Visa, MasterCard and American Express. These cards were issued by banks and financial institutions. Different types of credit cards categories are in a particular brand as well such as for low net worth, medium net worth and high net worth customers. To attract more customers, different incentives are offering such as airline miles, hotel room booking, restaurant dine-in, super market grocery buying, gift certificates to major retailers and cash back on purchases. Furthermore, in some banks have established rewards system for credit card usage. At the end of year these rewards points can be redeemed.

Branded versions of credit cards are issued to generate customer loyalty with store's name/ organization name emblazoned on the face of the cards. These credit cards called co-branded credit cards.

1.5.5 Credit Card Issuing Process

Before providing a credit card to the customer there is a process to establish a relationship with customer and the bank. Applying for a credit card for first time can be time consuming. Filling out an application form is mandatory and most bank nowadays allow to apply online by filling an application form. Choosing of suitable card can be done after self-studying or consulting sales executives. Figure 1.2 illustrated credit card issuing process as below.

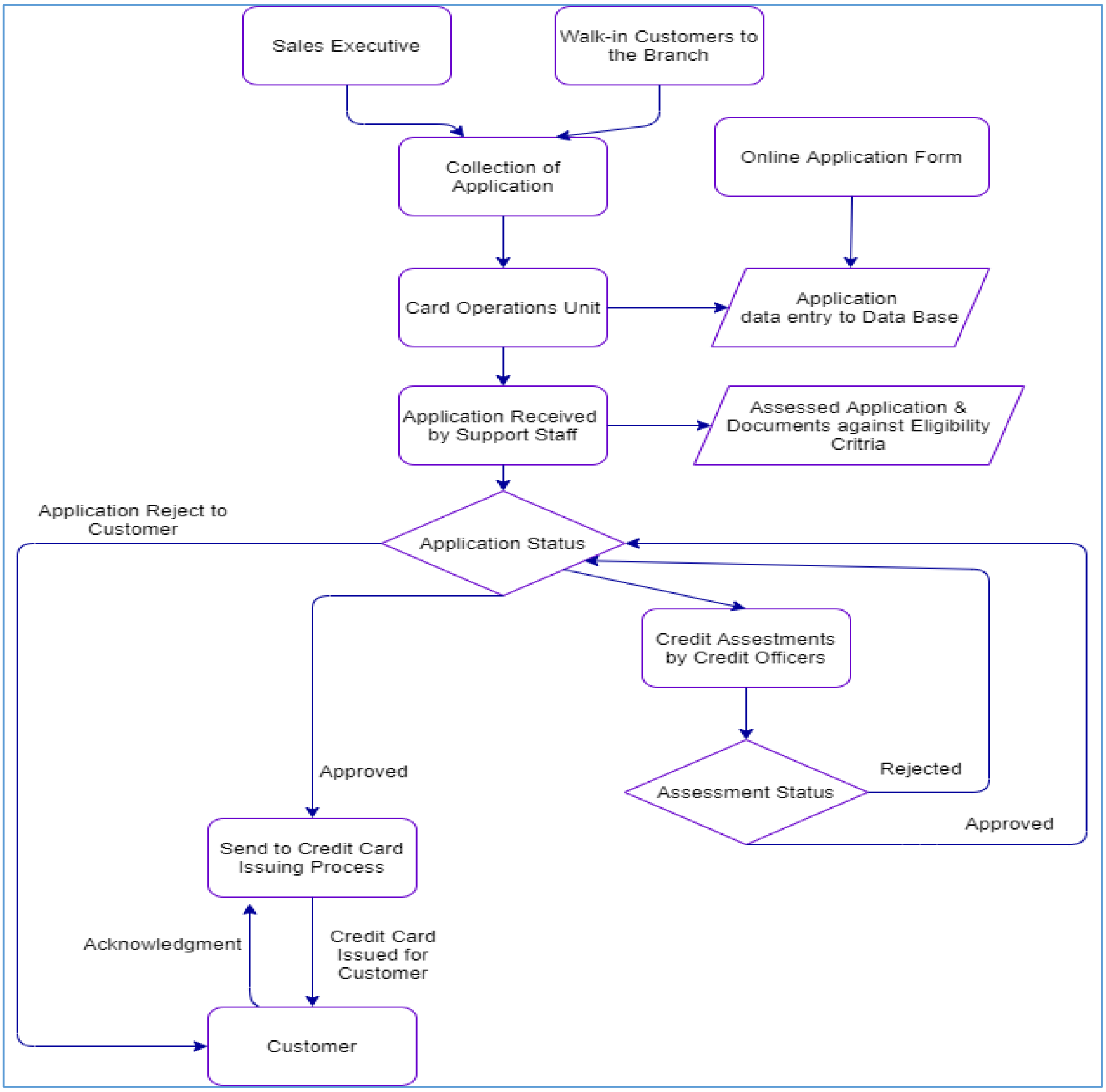

Sales Executive
Walk-in Customers to the Branch
Collection of Application
Online Application Form
Card Operations Unit
Application data entry to Data Base
Application Received by Support Staff
Assessed Application & Documents against Eligibility Critria
Application Reject to Customer
Application Status
Credit Assestments by Credit Officers
Approved
Assessment Status
Rejected
Approved
Send to Credit Card Issuing Process
Acknowledgment
Credit Card Issued for Customer
Customer

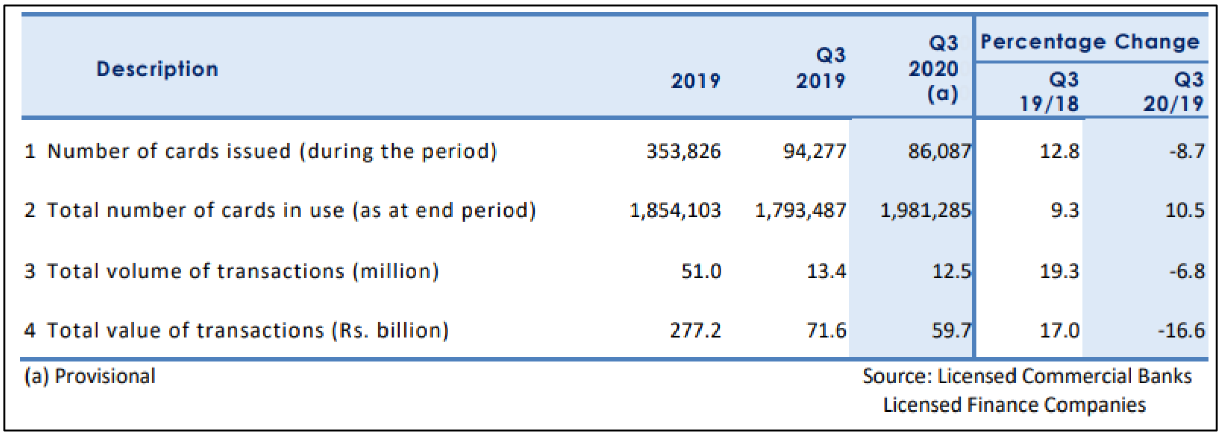
Figure 1.2 - Credit Card Issuing Process

Credit card application form with required supportive documents are handover by sales executive or walking customer to the branch. All applications will be handover to credit card operations unit. Application data entered to the Card Application Database. Some of bank offers online applications facility. These application data are being directly entered to the card application database by the applicant using the given online portal by the bank. Then the application review team assess & verify application, documents and eligibility criteria according to the internal policies and procedures. None eligible applications will be rejected to the customer. Eligible application is sent to credit assessment and by credit officers application approved with defining a LOC or rejected. Rejected application statuses will be communicated to the applicant. Approved applications will be added to the credit card issuing process. Eventually, applicant will receive a credit card and send an acknowledgement to the bank upon receiving and activating it.

1.5.6 Credit Card Business & Credit Card Defaulter Statistics of Sri Lanka

(Payments and Settlements Department, 2020) mentioned that issuing of credit cards started in 1989 by Commercial banks of Sri Lanka. To make more efficient the operation of credit card business, The Credit Card Operational Guidelines No. 1 of 2010 was issued. There are 14 Licensed Commercial Banks and 3 Finance Companies were licensed to engage in credit card business by the end of second quarter of 2020. Table 1.1 describe Credit Card statistics published by Central Bank of Sri Lanka.

Table 1.1 – Sri Lanka Credit Card Statistics *(Payments and Settlements Department, 2020)*



Description	2019	Q3 2019	Q3 2020 (a)	Percentage Change	
				Q3 19/18	Q3 20/19
1 Number of cards issued (during the period)	353,826	94,277	86,087	12.8	-8.7
2 Total number of cards in use (as at end period)	1,854,103	1,793,487	1,981,285	9.3	10.5
3 Total volume of transactions (million)	51.0	13.4	12.5	19.3	-6.8
4 Total value of transactions (Rs. billion)	277.2	71.6	59.7	17.0	-16.6

(a) Provisional

Source: Licensed Commercial Banks
Licensed Finance Companies

According to the above table 1.1 we can see that compared to 2019 third quarter there is growth in credit card usage in third quarter of 2020. However, there is decrease in credit card issuance and value of transaction in second quarter of 2020 this might be due to pandemic situation of COVID-19.

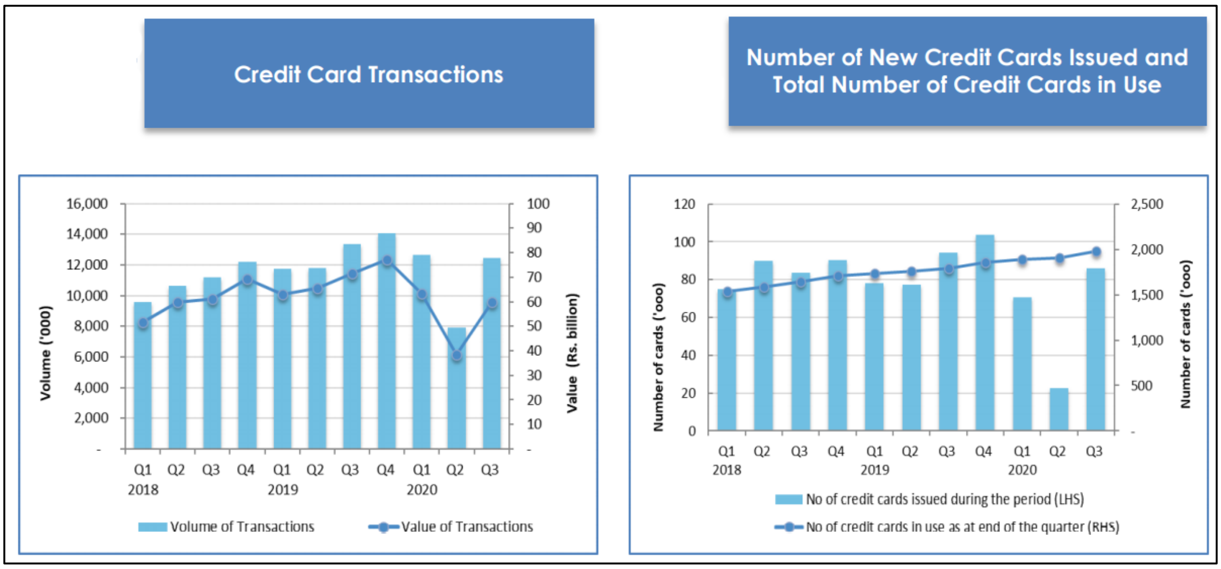


Figure 1.3 - Credit Card Transactions and Credit Cards Usage *(Payments and Settlements Department, 2020)*

Above figure 1.3 shows that there is a growth in credit card transaction in volume and value wise from 2018 to 2019. However, with the covid-19 situation there is a drop in second quarter of 2020. Credit card usage is increase from first quarter of 2018 to the end of the fourth quarter of 2020. New card issuance has an increasing trend in 2018 compared to 2021. However, there is drop in new card issuance in 2020 with COVID-19 effects. In general, 2019 second quarter is having effects on "Pasku" bombard terrorist attack situation which we can see slightly fluctuation in and 2020 quarter 1, 2 3 have impact on pandemic situation.

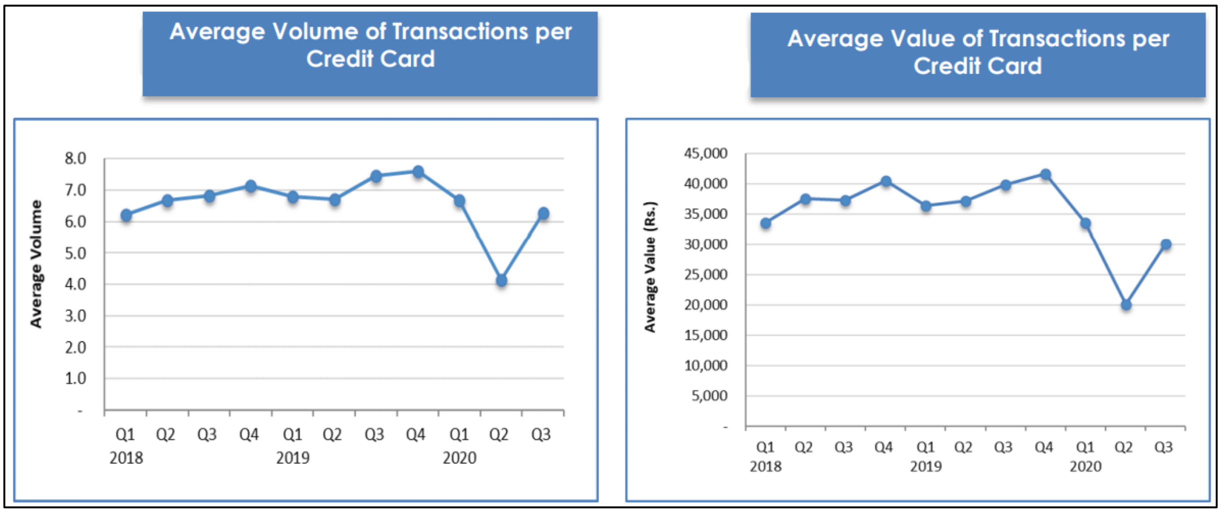
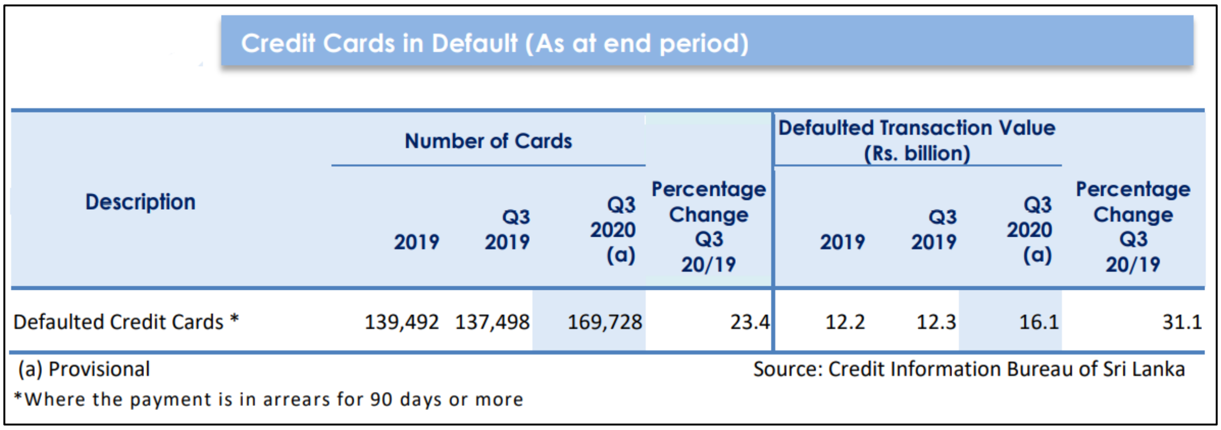


Figure 1.4 - Average Volume and Value of Transactions per Credit Card *(Payments and Settlements Department, 2020)*

According to the figure 1.4 average volume of transactions per credit card was increased in 2019 compared to 2018. However, there is a drop in 2020 with pandemic situation. Average value of transactions per card was increased in 2019 and decreased in 2020 with COVID-19 effects.

Table 1.2 - Credit Card Defaulter Statistics *(Payments and Settlements Department, 2020)*



Credit Cards in Default (As at end period)

Description	Number of Cards			Percentage Change Q3 20/19	Defaulted Transaction Value (Rs. billion)			Percentage Change Q3 20/19
	2019	Q3 2019	Q3 2020 (a)		2019	Q3 2019	Q3 2020 (a)	
Defaulted Credit Cards *	139,492	137,498	169,728	23.4	12.2	12.3	16.1	31.1

(a) Provisional

Source: Credit Information Bureau of Sri Lanka

*Where the payment is in arrears for 90 days or more

Table 1.2 shows credit card defaulter statistics. When it comes to the third quarter of 2020, credit card defaulter is increased in card and transaction both with COVID-19 effects. Percentage of change is 31 % which is very high.

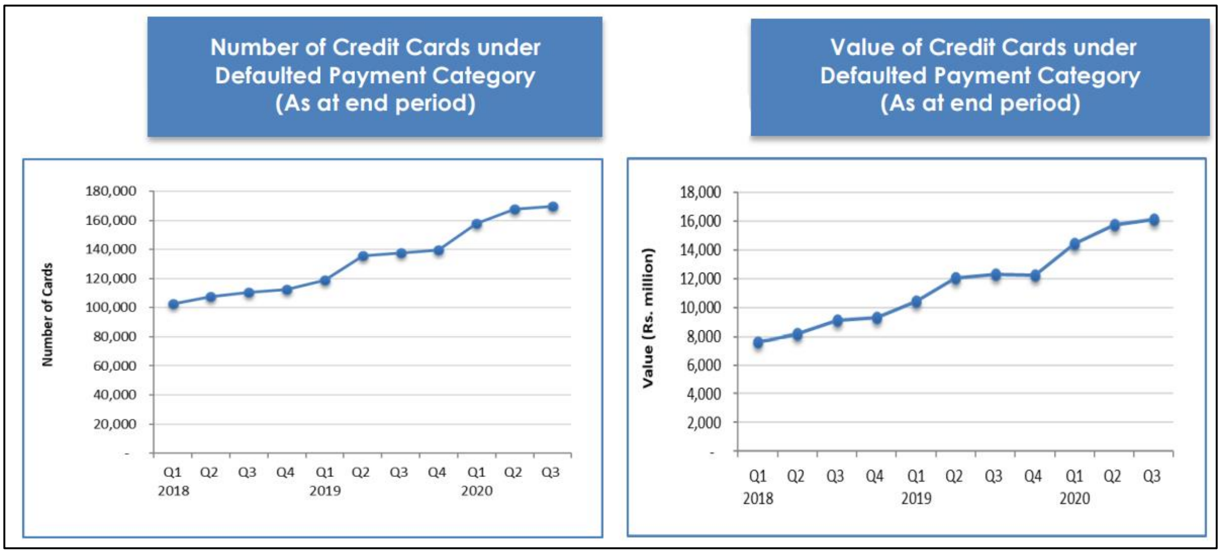


Figure 1.5 – No of Default Credit Card and Value *(Payments and Settlements Department, 2020)*

Figure 1.5 shows the number of defaulted credit cards and volume both increased in the third quarter of 2020. There is an increasing trend from 2018 to 2020 and this shows there is a need to decrease defaulters with which we can apply machine learning technologies.

1.6 Scope

The model will be developed and tested for the selected data source only. Artificial Neural Network (ANN) and Support Vector Mechanism (SVM) will be used to develop the model. Statistical methods under filter-based feature selection methods will be applied for feature selection. To check the accuracy of these models, Confusion Matrix, Area Under Curve (AUC) will be applied to the test data set. Finally evaluate ANN and SVM, will be used to compare the accuracy of two models and identify the best classification algorithm to predict customer eligibility for a credit card.

List of Deliverables

- Two ML models to predict credit card eligibility and its comparison
- Descriptive Statistics about data set
- Predicted data in a CSV file

1.7 Assumption and Limitation

Assumption of the project as follows.

- The model will be developed and tested for the selected data source and according to data / attributes availability only.
- To generate class variable from payment history data; payments are past due from 60th day and above consider as bad customers.

Limitation of the project as follows.

- The data set is taken from www.kaggle.com. Therefore, this data set might not be equaling to Sri Lankan Context (Sri Lankan customer’s behavior).
- COVID-19 effects on the data set is not available as the data set was built before covid-19 situation.

1.8 Structure of the Thesis

This report consists of five chapters. Chapter 01 delivered the brief introduction of this project to the reader including the background, authors’ perspective of the problem domain. Chapter 02 of this report is about literature review and it gives the related work and core concept behind the project. Chapter 03 provides the detailed information about the application of methodology which includes data pre and post processing, application of machine learning algorithms. Chapter 04 includes implementation of methodology including model validation and evaluation of models accuracy. Finally, Chapter 05 highlights the main findings and conclusion of the project.

CHAPTER 2

LITERATURE REVIEW

2.1 Risk Management at Banks

Saunders and Cornett (2014) states that the bank management's intendent to increase returns for shareholders which come with increased in risk. There are many risk faced by the banks. They are credit risk, interest rate risk, market risk, off-balance-sheet risk, technology and operational risk, foreign exchange risk, country or sovereign risk, liquidity risk, liquidity risk and insolvency risk. By effectively managing these risks, the banks can perform better. Furthermore, these risks and subject to regulatory attention due to the major role playing by banks in financial system. In (Leo et al., 2019) mentioned that "credit can be defined as the risk of potential loss to the bank if a borrower fails to meet its obligations (interest, principal amounts)". Credit risk is the single largest risk banks face.

2.2 Machine Learning & AI Implementation on Risk Management

Bhatore et al., (2020) carried out a comprehensive literature review on currently available research methods and machine learning techniques for credit risk evaluation. They have selected three major factors that create credit risk, careful examination and inspections while giving loans (credit scoring), continuous monitoring of customer payments and any other behavior patterns to decrease the probability of generating frauds (fraud detection) and non-performing assets (NPA). Further they have analyzed model evaluation techniques, current studies and research trends. Team have reviewed a total of 136 papers published between 1993 and March 2019 and concluded that Ensemble and Hybrid models with neural networks and SVM are more adaptive and mentioned that lack of complete public datasets will be cause for concern for researcher. Following figure summarized their findings about application of different ML techniques.

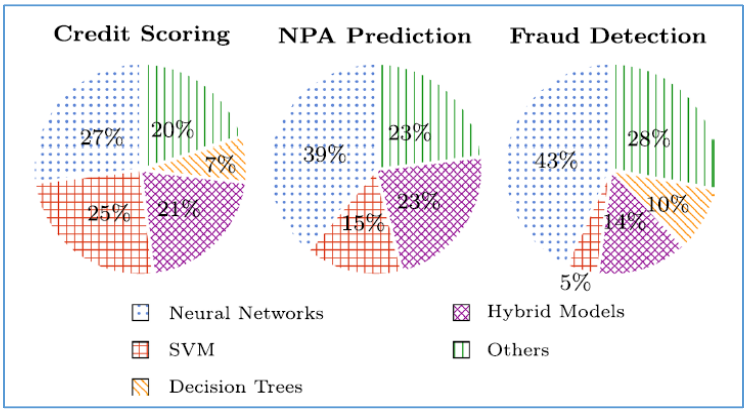


Figure 2.1 - ML techniques for Credit Risk Evaluation *(Bhatore et al., 2020)*

Leo et al., (2019) have been conducted a Literature Review on Machine Learning in Banking Risk Management. According to them (Bellotti and Crook, 2009; Huang et al., 2007; Li et al., 2017, Harris 2013) use SVM to develop scoring model for consumer credit management. Further they have mentioned (Yeh and Lien 2009; Galindo and Tamayo 2000; Keramati and Yousefi 2011) use Neural Network to build credit scoring model.

Banasik et al., (1999) mentioned that credit scoring systems were built to answer what likelihood of applicant of the credit facility to be received will be default in the future. Different modeling techniques use previous customers credit details and classified the customer as ‘good and ‘bad ‘considering their payment settlement pattern over a specified period. Furthermore, SriLaxmi et al., (2020) states that there are multiple criteria and factors considered when approving of a credit card. Mainly demographic, income, credit bureau data of the customer. Using credit card customer’s past data can identify key factors affect in credit risk by using models such as Logistic Regression and Random Forest. As a methodology they have Cross-Industry Standard Process for Data Mining (CRISP-DM). Moreover, Sariannidis et al., (2019) modeled seven classification methods, KNN, Logistic Regression, Naïve Bayes, Decision Trees, Random Forest, SVC, and Linear SVC and compared the prediction accuracy of these models. They have stated that in terms of lending decisions except few, most of the characteristic variables used can satisfactorily analyze default features. Additionally, it is important to have a better understanding of borrowers’ behavior with accounting, demographic and historical characteristics.

Karthiban et al., (2019) proposed a hybrid model which includes a novel 16-layer genetic cascade ensemble of classifiers, normalization techniques and two types of SVM classifiers. He used kernel functions, parameter optimizations, and stratified 10-fold cross-validation for feature extraction methods. The model achieved 97.39% prediction accuracy and concluded that proposed method can be applying in the banking domain to assess the bank credits of the applicants and aid the decision making process. Furthermore, Pristyanto et al., (2019) applied information gain, gain ratio, and correlation based feature selection (CBFS) and as a classifier used K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Artificial Neural Network. He concluded that feature selection does not always advance classifier accuracy but be subject to the characteristics and algorithms. Moreover, Munkhdalai et al., (2019) performed a broad comparison between the machine-learning approaches and a human expert-based model FICO credit scoring system by using a Survey of Consumer Finances (SCF) data. To reduce the computation cost and to choose the most informative variables they have applied two variable-selection methods for feature-selection. In this study, they present TSFFS (two-stage filter feature selection) algorithm and use the NAP method for variable-selection. As ML techniques logistic regression, support vector machines, an ensemble of gradient boosted trees and deep neural networks used. They have concluded subset selected by NAP from the deep neural networks and XGBoost algorithms trained on achieve the very best accuracy and area under the curve (AUC).

Comparative evaluation of the performances of five popular classifiers namely, Naive Bayesian Model, Logistic Regression Analysis, Random Forest, Decision Tree, and K-Nearest Neighbor Classifier use in credit scoring used to carried out by (Wang et al., 2020). They have concluded that each individual classifier has its own strength and weakness. However, the results of this experiment discover that Random Forest achieves better results than others in terms of precision, recall, AUC (area under curve) and accuracy. However, Karthiban et al., (2019) applied the Regression, Naive Bayes, Generalized Linear model, Deep learning(DL), Decision tree, Random Forest and Gradient Boosted trees were for Bank Loan Approval data set. They used confusion matrix to evaluate models. In there they have consider Accuracy, Sensitivity or True Positive Rate or Recall, Specificity or True Negative Rate, Precision, F measure, Classification Error, AUC, ROC Curve for evaluation. Moreover, Antonakis and Sfakianakis (2009) benchmark two data sets NBR against linear discriminant analysis, logistic regression analysis, k-nearest neighbors, classification trees and neural networks. He concluded that considering all measures used, NBR is found to have lower predictive power than other five classifiers in each data set.

To accurately identify loan defaulters (Shoumo et al., 2019) applied support vector machine, extreme gradient boosting, logistic regression and random forest classifiers to a loan data set. Dimensionality reduction carried out by using Recursive Feature Elimination with Cross Validation and Principal Component Analysis. To model evaluation metrics such as F1 score, AUC score, prediction accuracy, precision and recall have been used. They have concluded that support vector machines can outperform other tree-based models or regression models. Furthermore, they have concluded that the model has shown that recursive feature elimination with cross-validation can outperform models based on principal component analysis. However, Agarwal et al., (2020) use different classification algorithms were evaluated such as Logistic Regression, Decision Tree, K-Nearest Neighbor and Naive Bayesian for credit card dataset. The dataset is obtained from UCI Repository credit card defaulter. Main objective to compare the performance measures between the original dataset and original dataset with the principal component is applied. Benchmarked before and after applying the principal component. Different algorithms are compared on the basis of various metrics such as Accuracy, Precision, F1-Score, Recall, ROC. He concluded that Logistic regression is performed better in this particular data set in accuracy and precision measures. The other performance measures such as ROC, F1-Score showed good results for naïve Bayesian. K-Nearest neighbor showed acceptable performance in terms of recall.

Kumar Gupta and Goyal (2018) applied Artificial Neural Network (ANN) to predict the creditworthiness of an application. Data set has been taken from kaggle.com (lending club loan data). Dependent Variable is loan status (0 and 1). Scoring system develop by using discriminant analysis. They have concluded that results of both the systems have shown an equal outcome on the dataset. The classifier is very effective with the accuracy of 97.68% in artificial neural network. The system classifies the predicted variable correctly with a very low error. Hence, both models can be used to identify credit default with equal accuracy. However, Lee and Chen (2005) evaluate the performance of credit scoring using two-stage hybrid modeling methods with artificial neural networks and multivariate adaptive regression splines (MARS). They have used n-fold cross-validation to reduce the possible bias linked with the random sampling of the training and testing samples and the entire dataset is randomly split into mutually exclusive n numbers. To build the two-stage hybrid model, a single-layer BPN model again applied. Important independent variables gained from the MARS were input to the input layer of the hybrid model. He concluded that the proposed hybrid method outperforms the results using discriminant analysis, logistic regression, artificial neural networks and MARS.

Wang et al., (2011) Carried out comparative assessment of the performance of three ensemble methods; Bagging, Boosting, and Stacking with four classifiers namely Logistic Regression Analysis (LRA), Decision Tree (DT), Artificial Neural Network (ANN) and Support Vector Machine (SVM). They have discovered that the three ensemble methods can significantly improve individual base learners. Precisely Bagging performs better than Boosting across all credit datasets. In terms of average accuracy, type I error and type II error; Stacking and Bagging DT get the best performance in their experiments. Furthermore, Marqués et al., (2012) use two resampling-based ensembles (bagging and AdaBoost) and two attribute-based algorithms (random subspace and rotation forest) in various sequences. To compare the performance of the rotation forests with other classifier ensembles six real-world credit data sets used. Fivefold cross-validation method has been adopted and to evaluate accuracy, error rate, Gini coefficient, Kolmogorov– Smirnov statistic, mean squared error, area under the ROC curve, and Type-I and type-II errors used. Their experimental results and statistical tests disclosed that new two-level classifier ensemble based approaches are a suitable solution for credit scoring problems performing better than the traditional single ensembles and individual classifiers.

Chornous and Nikolskyi (2018) proposed an ensemble-based classification model with business related feature selection to increase accuracy of classification of credit scoring. The data set was collected from Vidhya loan prediction hackathon and contains of 614 observations. He has selected Information Gain, Chi-Squared and Mean Decrease Gini as feature selection methods. He concluded that a hybrid approach for user-defined variables can be more effective in ensemble binary classification models. Furthermore, Oreski et al., (2012) propose a feature selection technique for finding an optimum feature subset which makes neural network classifiers high in accuracy. The feature selection techniques used here is Genetic algorithm, Forward selection, Information gain, Gain ratio, Gini index and Correlation. Credit dataset collected at a Croatian bank used to conduct the experiment. They have concluded that discovering the most important features in defining the risk of a default, hybrid system with a genetic algorithm can be used as feature selection techniques.

Madyatmadja and Aryuni (2005) study to discover an appropriate data mining method for credit scoring credit card application in a Bank and improve the performance. Their proposed model of classification applies Naïve Bayes and the ID3 algorithm. The class variable in the data set is classified into two class labels as ‘approve’ and ‘reject’. By using the credit experts’ knowledge, the class label is determined. They have got 82% Accuracy on Naïve Bayes classifier and 76% accuracy on ID3. They have concluded that the Naïve Bayes classifier performed better with high accuracy than the ID3 classifier. Furthermore, Hamid and Ahmed (2016) build a new model for categorizing loan risk in the banking sector by using data mining to predict the status of loans. Three algorithms have been used to build the proposed model: j48, bayesNet and Naïve Bayes. The developments were carried out with Weka application. They have concluded that J48 was selected as the best algorithm based on its high accuracy and low mean absolute error as shown in the result.

Blagus and Lusa (2013) studied the behavior of Synthetic Minority Oversampling Technique (SMOTE) for high-dimensional class-imbalanced data sets. They have concluded that SMOTE is very efficient in low dimensional data sets and less effective on high dimensional data. If feature selection is performed before application of SMOTE beneficial for k-NN classifiers when data are high in dimensional. Furthermore, (Elreedy and Atiya, 2019) mentioned the accuracy of SMORTE for generally declines with higher dimension. If number of minority examples N is high, accuracy can be increase.

CHAPTER 3

METHODOLOGY

Systematic Approach

To carry out the project, CRISP-DM frame work was used as shown in Figure 3.1 and detail discussion of each phase relevant to application for project is listed below.

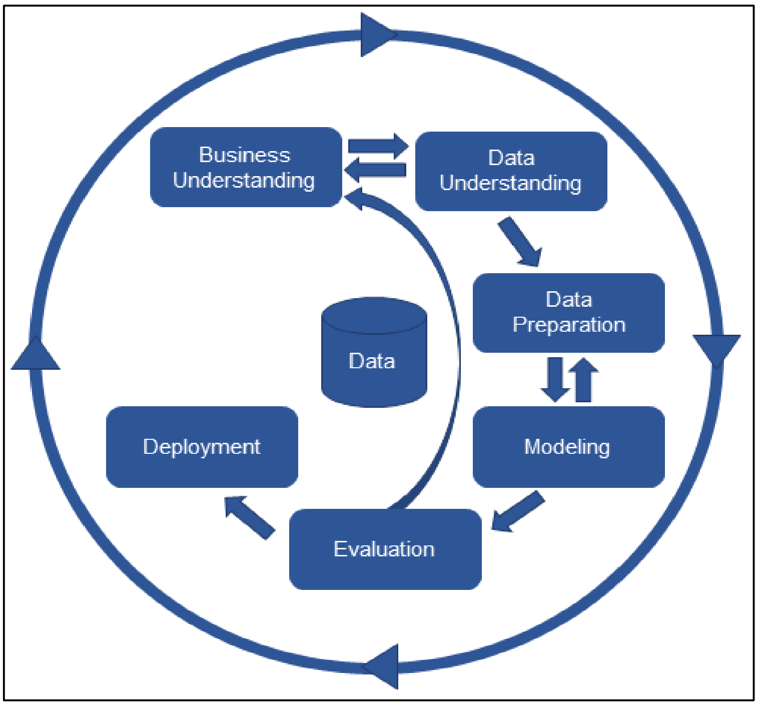


Figure 3.1 - CRIP –DM Model *(Taylor, 2017)*

CRISP-DM (Cross-industry standard process for data mining) data mining process was published in 1999 to standardize. There 6 phases, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and deployment. Brief description of each phases are listed below (Data Science Process Alliance, n.d.).

- **Business Understanding** - Understanding of objectives and requirements and produce of detail plan for project focus in here.
- **Data Understanding** - Focusing on identify, collect and analyze data. Data format/fields identification, identify relationships by visualization; verify data quality (clean/dirty) are the main activities carried during this phase.

- **Data Preparation** – This phase often called ‘data munging or wrangling. Selection of data, clean data, construct data, integrated data and format data are basic activities carried out under this phase.
- **Modeling** – Determine selection of algorithms, generate test design, build model and asses model are main activities carried out in this phase.
- **Evaluation** – Focusing on identification of which model best fit the business requirement. Evaluate results, review process, determine next step are key activities in here. By determining whether to proceed to deployment or iterate further will be judge in here.
- **Deployment** – Focusing on accessible methods for developed model output/results. Deployment plan, monitoring and maintenance, produce final report and review project are key activities in here.

3.1 Business Understanding

Credit card is one of the key lending product facilities given for a customer by a bank. The repayments of credit card are always not guaranteed and it often ends up as non-performing credit facility (NPL). Banks are assessing the background check of the individual customers by analyzing their eligibility, yet the bank sometime end up in making wrong selections. The credit card has a higher annual percentage rate (APR) and by law, card issuers must provide 21 days of grace period before interest on purchases and begin to accrue. When customers paying the balance before the grace period expired consider as a good practice. For any unpaid balance normally after one month of purchase is made Interest charges will apply. Any un paid balance carried forward from previous month and for new charges grace period will not be provided. According to country’s financial policy interest will be accruing daily or monthly.

3.2 Data Understanding

The data set has been taken from kaggle.com data repository (Song, 2019). This data set is publically available data set. Hence information security is not a concern in here.

URL - https://www.kaggle.com/rikdifos/credit-card-approval-prediction/tasks?taskId=1416

There are two data set and detail of data shown in Table 3.1 and 3.2.

- application_record.csv for applicant information – No of records 438,510
- credit_record.csv for credit record information – No of records 1,048,576

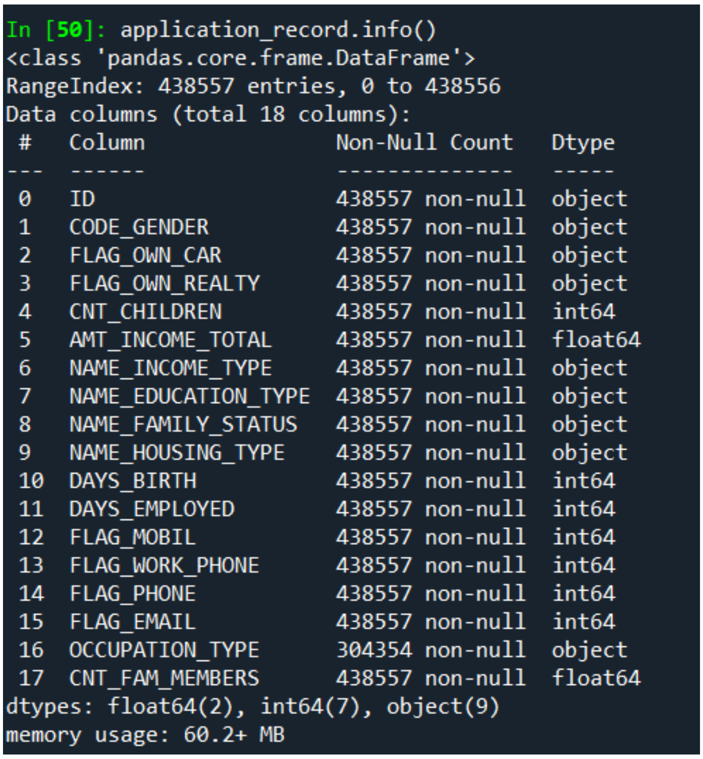
Table 3.1- Detail Information about application data set

Credit Card Application Data			
Feature Name	**Explanation**	**Data Type**	**Possible Values**
ID	Client number	Numerical	
CODE_GENDER	Gender of the client	Categorical	M, F
FLAG_OWN_CAR	Is there a car	Categorical	N , Y
FLAG_OWN_REALTY	Is there a property	Categorical	N, Y
CNT_CHILDREN	Number of children	Numerical - Integer	
AMT_INCOME_TOTAL	Annual income	Numerical - float	
NAME_INCOME_TYPE	Income category	Categorical	Commercial associate Pensioner, State servant Student ,Working
NAME_EDUCATION_TYPE	Education level	Categorical	Academic degree, Higher education, Incomplete higher, Lower secondary, Secondary / secondary special
NAME_FAMILY_STATUS	Marital status	Categorical	Civil marriage, Married, Separated, Widow Single / not married,
NAME_HOUSING_TYPE	Way of living	Categorical	Co-op apartment, House / apartment, Municipal apartment, Office apartment Rented apartment, With parents
DAYS_BIRTH	Birthday	Numerical - Integer	Count backwards from current day (0), -1 means yesterday
DAYS_EMPLOYED	Start date of employment	Numerical - Integer	Count backwards from current day (0). If positive, it means the person currently unemployed.
FLAG_MOBIL	Is there a mobile phone	Numerical - Integer	1 , 0
FLAG_WORK_PHONE	Is there a work phone	Numerical - Integer	1 , 0
FLAG_PHONE	Is there a phone	Numerical - Integer	1 , 0
FLAG_EMAIL	Is there an email	Numerical - Integer	1 , 0
OCCUPATION_TYPE	Occupation	Categorical	Several occupation
CNT_FAM_MEMBERS	Family size	Numerical - Float	

Table 3.2 - Detail Information about payment history data set

Credit Card Record (Payment History Data)			
Feature Name	**Explanation**	**Remarks**	
ID	Client number	Numerical - Integer	
MONTHS_BALANCE	Record month	Numerical - Integer	The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on
STATUS	Status	Categorical	0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month

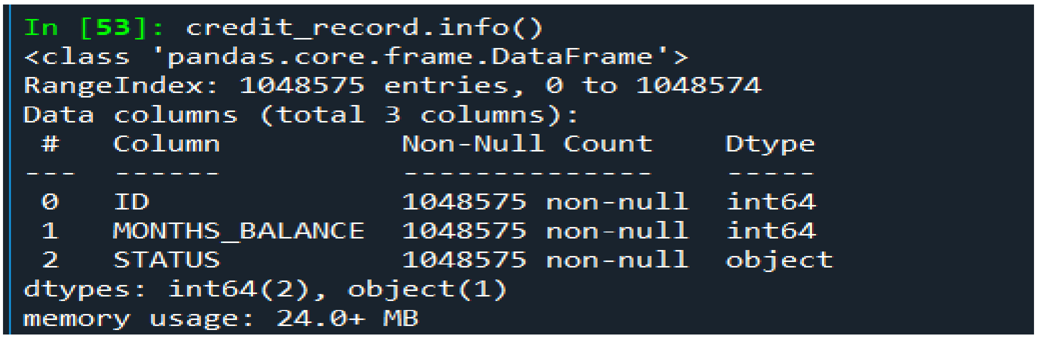
Application record file has nine categorical variables and nine numerical variables as shown in figure 3.2. According to the figure there are null data in occupation_type column. The data set does not contain direct class variable.



```
In [50]: application_record.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 438557 entries, 0 to 438556
Data columns (total 18 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   ID                   438557 non-null  object
 1   CODE_GENDER          438557 non-null  object
 2   FLAG_OWN_CAR         438557 non-null  object
 3   FLAG_OWN_REALTY      438557 non-null  object
 4   CNT_CHILDREN         438557 non-null  int64
 5   AMT_INCOME_TOTAL     438557 non-null  float64
 6   NAME_INCOME_TYPE     438557 non-null  object
 7   NAME_EDUCATION_TYPE  438557 non-null  object
 8   NAME_FAMILY_STATUS   438557 non-null  object
 9   NAME_HOUSING_TYPE    438557 non-null  object
 10  DAYS_BIRTH           438557 non-null  int64
 11  DAYS_EMPLOYED        438557 non-null  int64
 12  FLAG_MOBIL           438557 non-null  int64
 13  FLAG_WORK_PHONE      438557 non-null  int64
 14  FLAG_PHONE           438557 non-null  int64
 15  FLAG_EMAIL           438557 non-null  int64
 16  OCCUPATION_TYPE      304354 non-null  object
 17  CNT_FAM_MEMBERS      438557 non-null  float64
dtypes: float64(2), int64(7), object(9)
memory usage: 60.2+ MB
```

Figure 3.2 - Information of Application record data

Credit card records has one categorical column and two numerical columns as shown in figure 3.3. This data set contain duplicate data for ID column.



```
In [53]: credit_record.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 3 columns):
 #   Column          Non-Null Count    Dtype 
---  ------          --------------    ----- 
 0   ID              1048575 non-null  int64 
 1   MONTHS_BALANCE  1048575 non-null  int64 
 2   STATUS          1048575 non-null  object
dtypes: int64(2), object(1)
memory usage: 24.0+ MB
```

Figure 3.3 - Information of credit record data

3.3 Data Preparation Methods

Data preparation phase, which is often referred to as “data munging” or “Data Preprocessing” prepares the final data set(s) for modeling. Python programing with libraries /packages use to prepare the data set.

Key main areas related to data preparation phase considered in the project as follows.

- Data Preparation with Explanatory Data Analysis (EDA) under each preparation activity
- Feature Selection from finally prepared data set

Figure 3.4 shows identified different data preparation activities related to our project and each activity will discuss separately below.

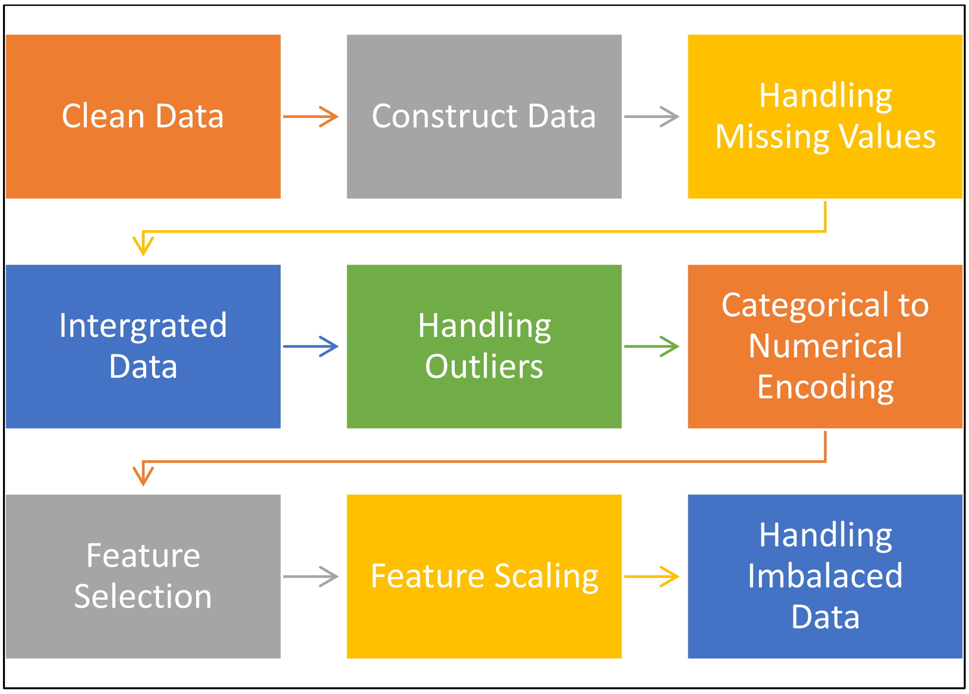


Figure 3.4 - Activities in Data Preparation Phase

3.3.1 Clean Data

Data set might contain erroneously entered data. These erroneous values need to correct, impute or removed from the data set.

3.3.2 Handling Missing Value

Missing value occurred may be due to many reasons. By handling missing values, it will increase performance of the model. Common methods are replacing missing values with mean or median of entire column (imputation) or deleting rows/ columns.

3.3.3 Construct Data

Derive new attributes from exiting data set.

3.3.4 Integrated Data

Integrating data phase basically combined data from multiple sources.

3.3.5 Outlier Removals

Outlier is a data point that differs significantly from other observations. To remove outliers, we can use statistical method Inter Quartile Range (IQR) and removed outliers from the data set. Final data set is 33,140. Figure 3.5 shown main component of IRQ.
below 25th percentile – 1.5 * IQR, or above 75th percentile + 1.5 * IQR

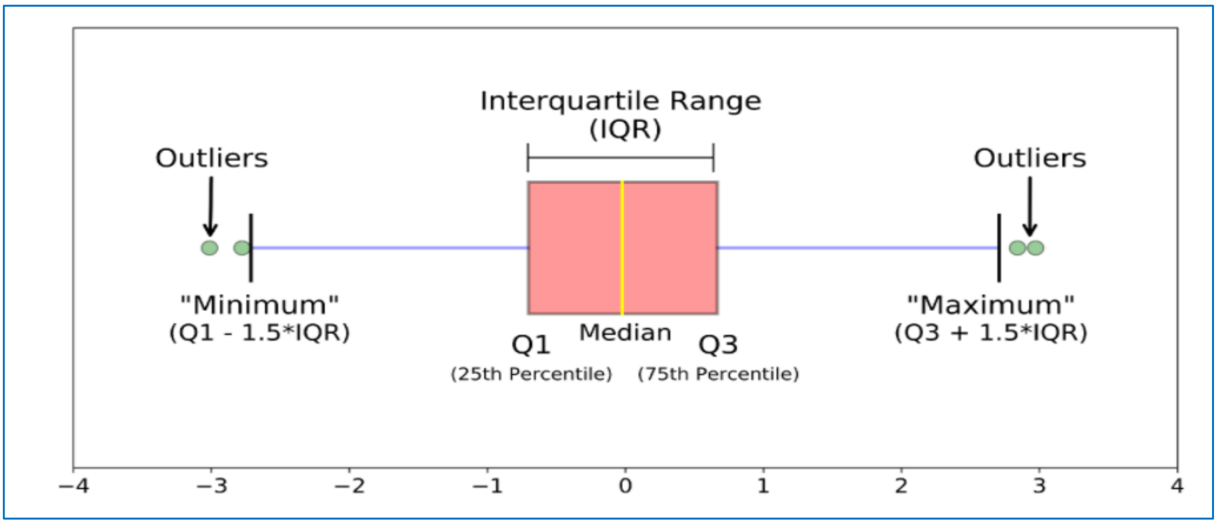


Figure 3.5 - Interquartile Range *(Galarnyk, 2018)*

3.3.6 Encoding Categorical Data

Categorical data can't be use at mathematical equations. Such as 'Male' and 'Female' in gender column. These columns need to convert to numerical values. There are varies methods can apply for categorical encoding by considering categorical feature is ordinal or nominal.

3.3.7 Feature Selection

Statistical method under filter-based feature selection applied for feature selection as shown in the figure 3.6.

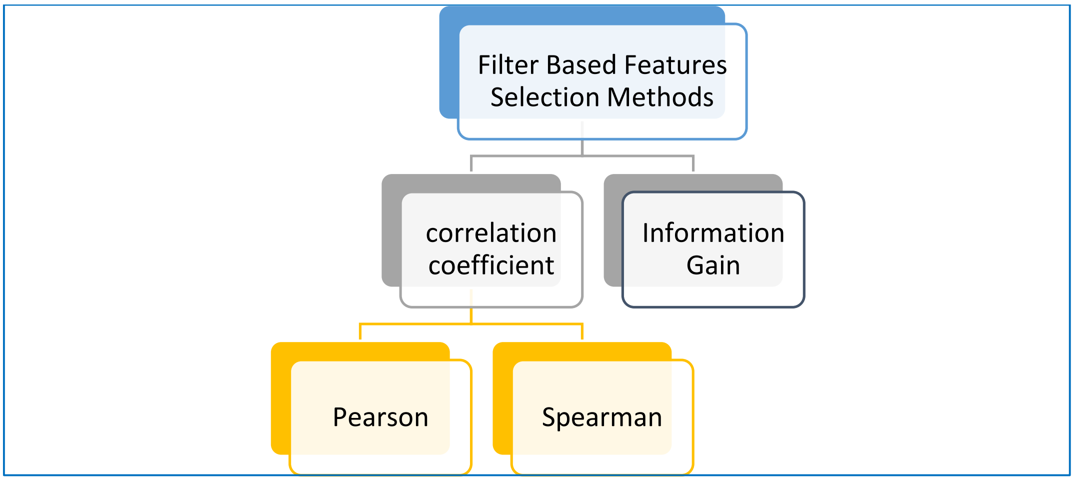


Figure 3.6 - Feature Selection Methods

Correlation Based Feature Selection

Correlation is a bi-variate investigation and it asset association between two variables and the way of the relationship. Correlation coefficient value varies between +1 and -1. A value of + 1 shows a perfect relationship between the two variables. Relationship of two variables will weaker when correlation coefficient value goes towards 0. The + sign indicates a positive relationship and a - sign indicates a negative relationship. Main difference between each correlation methods explained as follows (Statistics Solutions, n.d.)

Pearson Correlation Coefficient – To measure relationship between linearly related variables Pearson r correlation is the most widely used. Both variables should be normally distributed (normally distributed variables have a bell shape curve).

Spearman Rank Correlation - To measure the degree of association between two variables Spearman rank correlation is use. Spearman rank correlation is a non-parametric test. (Statistics Solutions, n.d.) mentioned that "It does not carry any assumptions regarding the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal”.

Information Gain

Information Gain sometimes denoted as Mutual Information measure the dependence between the two variables. It measures information value of each independent variable respect to dependent variable and select the one has most information gain. The variable considers as more dependent when the information gain value is high.

3.3.8 Feature Scaling

We have applied Standard scaler for numerical features. Standardization is highly used in SVM and ANN. In here transformed set of numerical values making mean equal to 0 and standard deviation equal to 1.

3.3.9 Handling Imbalance Data

Class imbalance is common problem in machine learning which is inherited by nature for default prediction, customer churning etc. Class imbalance is number of observation belong to one class is significantly lower than the other class.

To balanced data, we can use over sampling or under sampling. Both method might be cause model over fitting unless if we use correct technology.

- Oversampling happened when adding more copies of the minority class. Oversampling can be a correct choice when you have less data.
- Under sampling happened when removing some observations of the majority class. Under sampling can be a correct choice when you have large data set such as millions of rows. Figure 3.7 shown oversampling and under sampling.

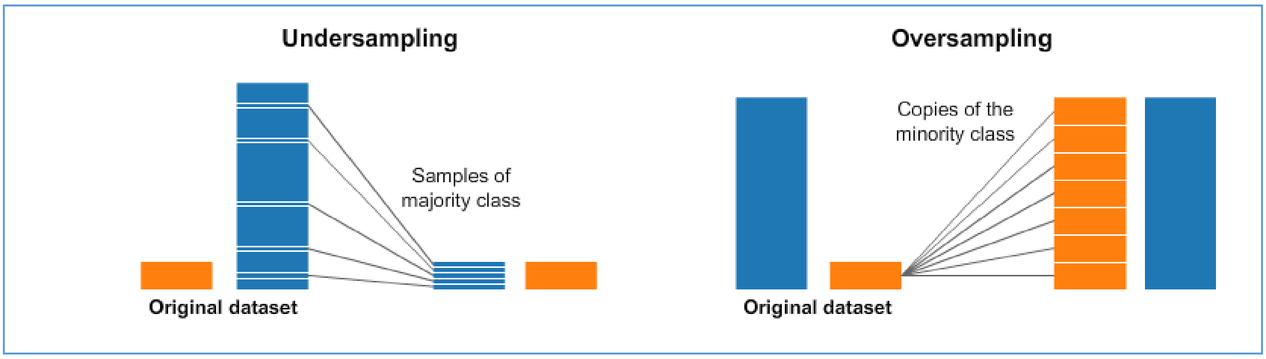


Figure 3.7 - Oversampling and Under Sampling *(Kumar, 2021)*

In here we have use synthetic sampling approach, SMOTE (Synthetic Minority Oversampling Technique) to handling imbalanced data. Our data set have 32,586 records and dimensions are not in very high. This technique produces synthetic data for the minority class as shown in figure 3.8. In SMOTE, randomly picking a point from minority class and compute k-nearest neighbors for selected point. Between selected point and its neighbor's synthetic points are added (Kumar, 2021). **SMOTE algorithm** works in 4 simple steps (Hussein et al., 2019):

1. Discover the k-nearest neighbors for each sample.
2. From a k-nearest neighbor select samples randomly
3. Find the new samples
4. Add new samples to the minority. Repeat the steps until data is balanced.

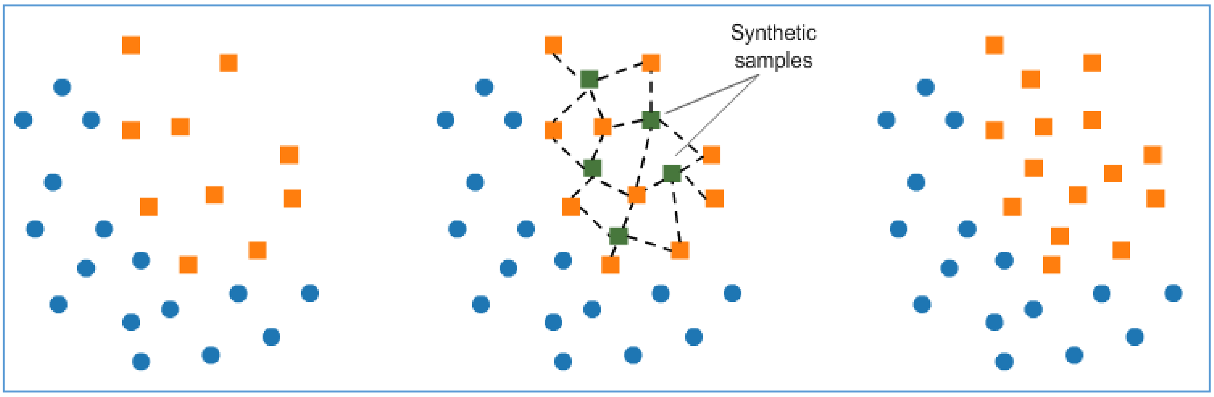


Figure 3.8- Synthetic Minority Oversampling Technique (SMOTE) *(Kumar, 2021)*

3.4 Modeling

We have acquired relevant data set and data preparation with feature selection was done and finalized our data set. Then applied standard scaler to numerical data for data scaling and apply SMORTE for finalized data set. Next step is to divide the data set as a training and test into a ratio of 80:20. Training data set is used to train the model by applying ANN and SVM. In here use linear and nonlinear SVM both models. Python programming and its libraries have been used to develop the models. Finally evaluate the predicted results of ANN and SVM, compare the accuracy of two models by using Mean Squared Error and Confusion Matrix to choose the most accurate model. Test data set used to test the model and evaluate the outcome. Workflow of the modeling process shown in figure 3.9.

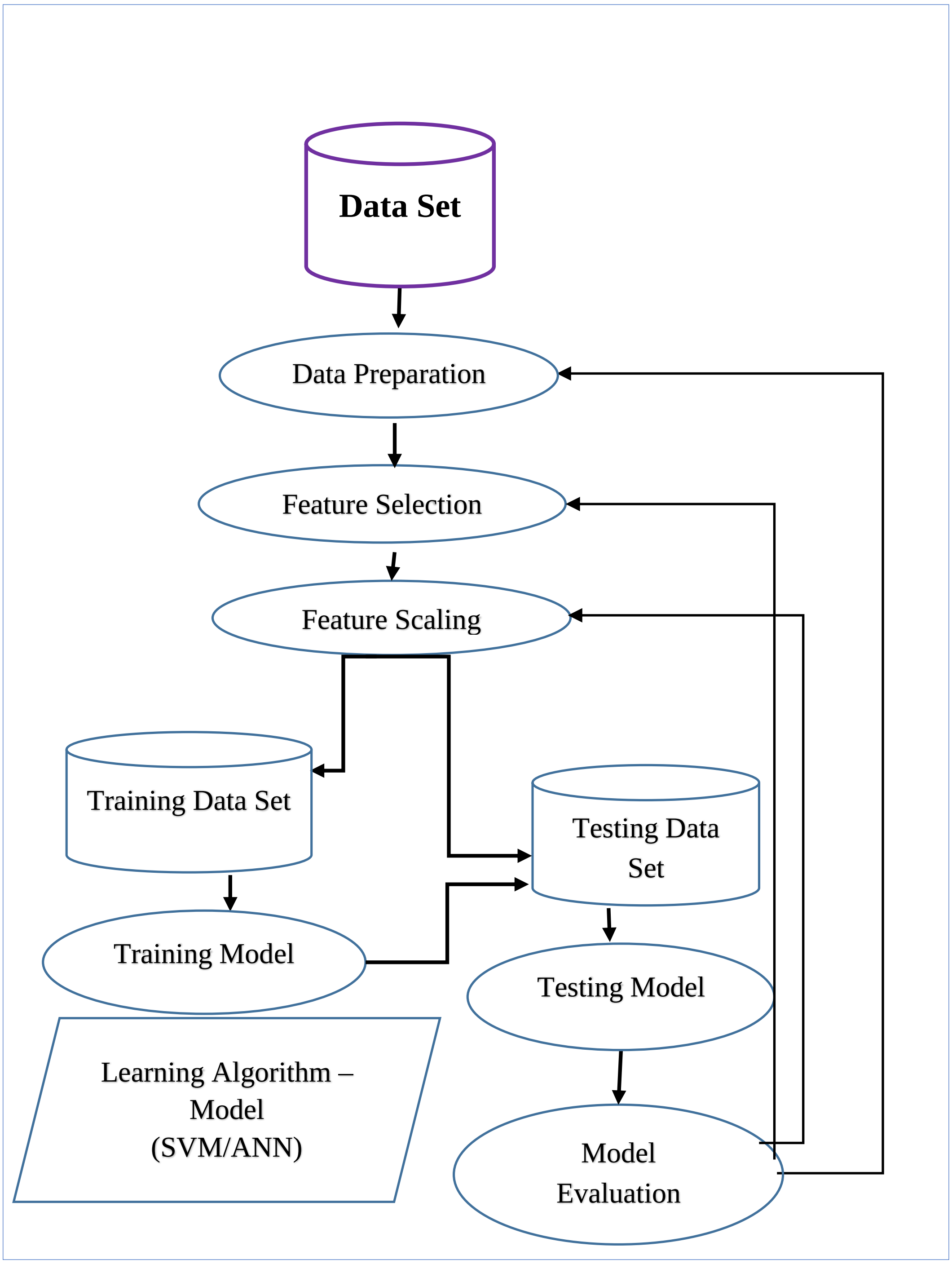


Figure 3.9 – Modeling Work Flow

3.4.1 Artificial Neural Network

Artificial Neural Network (ANN) is evolved from biological neural network of human brain. It is deep learning algorithm and use as information processing technique. We can use ANN not only for a classification problem but also regression. Neural network may contain 3 layers as follows:

- Input Layers – Raw information feed as input to the network
- Hidden Layer – Input unit and weight. There can be many hidden layers.
- Output Layer – This layer depends on hidden layer and weights or input layer. Prediction related to response variable return in output layer.

Perceptron is a single layer neural network and it is a linear classifier (binary) used in supervised learning. Perceptron helps to classify the given input data.

There are two important types of Artificial Neural Networks

- Feed-Forward Neural Network
- Feed-Back Neural Network

Feed-Forward Neural Network

The flow of information goes only one direction from input layer, hidden layer and finally to output layer in Feedforward networks. Feed forward ANNs is a supervised learning technique and commonly used in classification problems. In this study we are focusing on the feed-forward neural network. In ANN there are multiple and hyper parameters that affect the performance of the model and its output. Figure 3.10 shows layers of Feed-Forward Artificial

Neural Network.

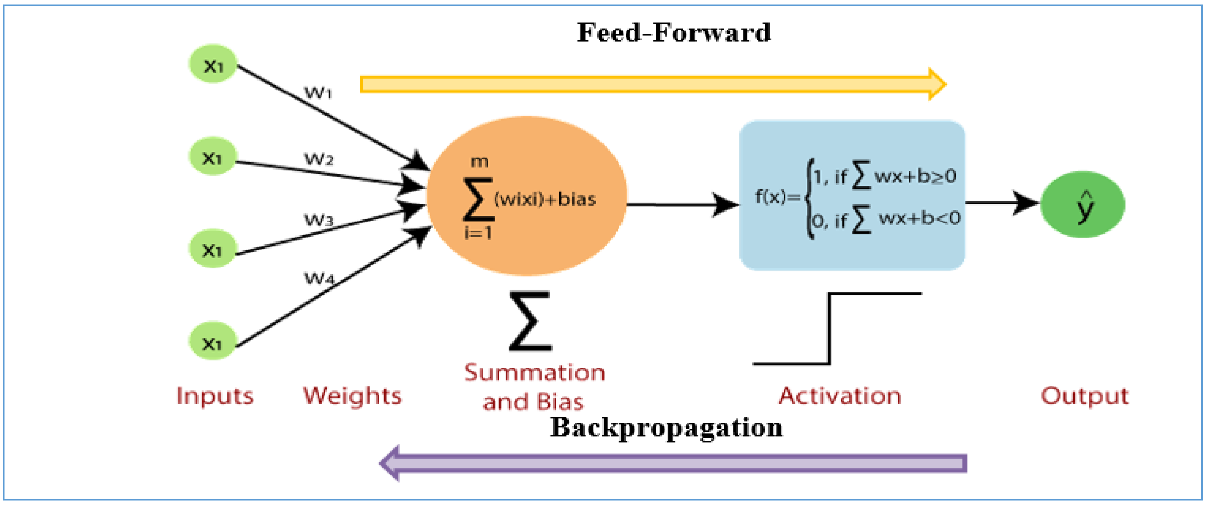


Figure 3.10 - Layers of Feed-Forward Artificial Neural Network *(Nellur, 2020)*

Steps Involve in application of ANN

1. Input – 'X' is input variable with some value and each input connected to hidden layer as shown in below figure 3.11

   x1, x2, x3, x4, x5 ….x21

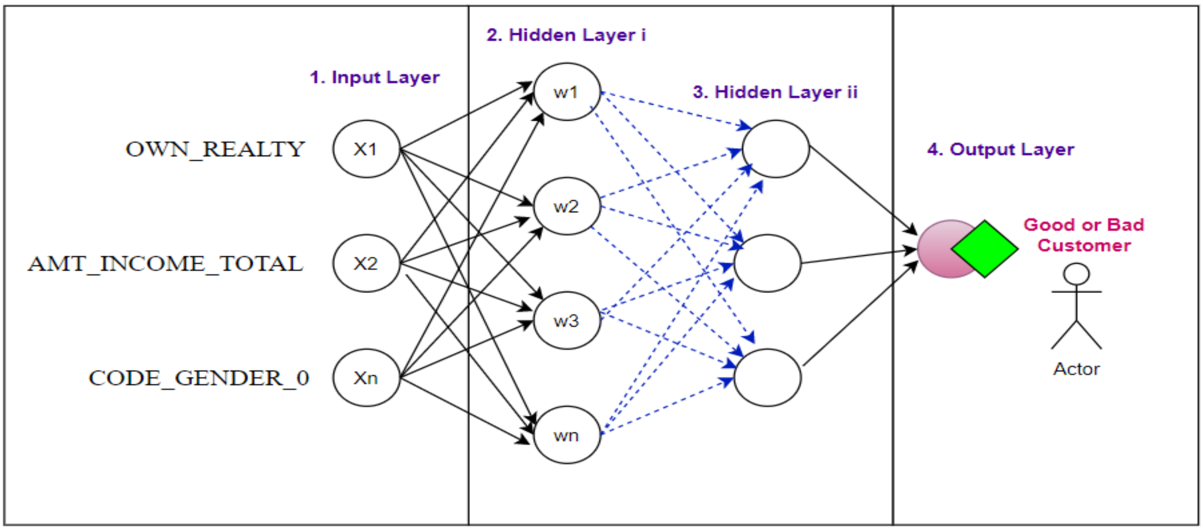


Figure 3.11 - ANN Architecture

2. Weight – Assigned weights for input. In the beginning weights are randomly initialized and later it will get adjusted consequently.

   w1, w2, w3, w4, w5, w6 .., wn
3. Summation - Summation of input with respective weights

   x1w1 + x2w2 + x3w3 + x4w4 + …. xnwn
4. Biasness – Biasness factor added, and it is a constant value. We can reduce over-fitting the data set by adding biasness.
5. Activation Function – Is used to get out put value Node should be activating or not is decides by the activate function depending on the sum of input variable with bias. The activation function is applied to hidden layer and output layer. Commonly use activation functions are Threshold Activation Function, Sigmoid Activation Function, Tangent Hyperbolic Activation Function, Rectifier Activation Function. In this project select sigmoid function as activate function as in figure 3.12.

Activation function = [x1w1 + x2w2 + x3w3 + x4w4 + …. xnwn] + bias (1)

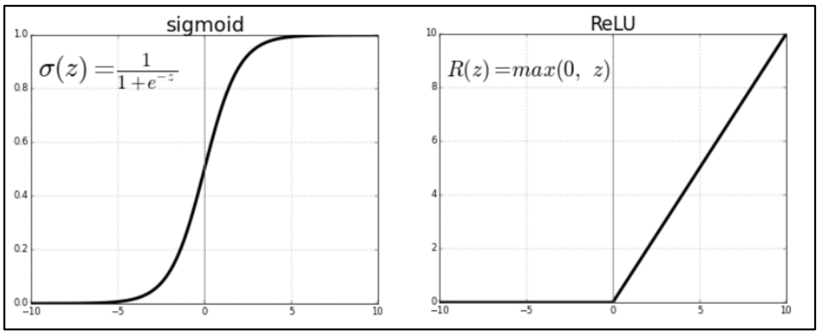


Figure 3.12 - Sigmoid and ReLU activate Functions *(Sharma, 2017)*

6. Application of Forward Propagation & Back Propagation in ANN

Forward Propagation: Data fed into input layer carried out to hidden layer by using activate function applied to the node and carried forward to output layer then to output the result.

Backpropagation is vice versa of forward propagation. It is coming backward from output value, hidden layers to adjust weights to minimize error in output value as shown in the figure 3.13.

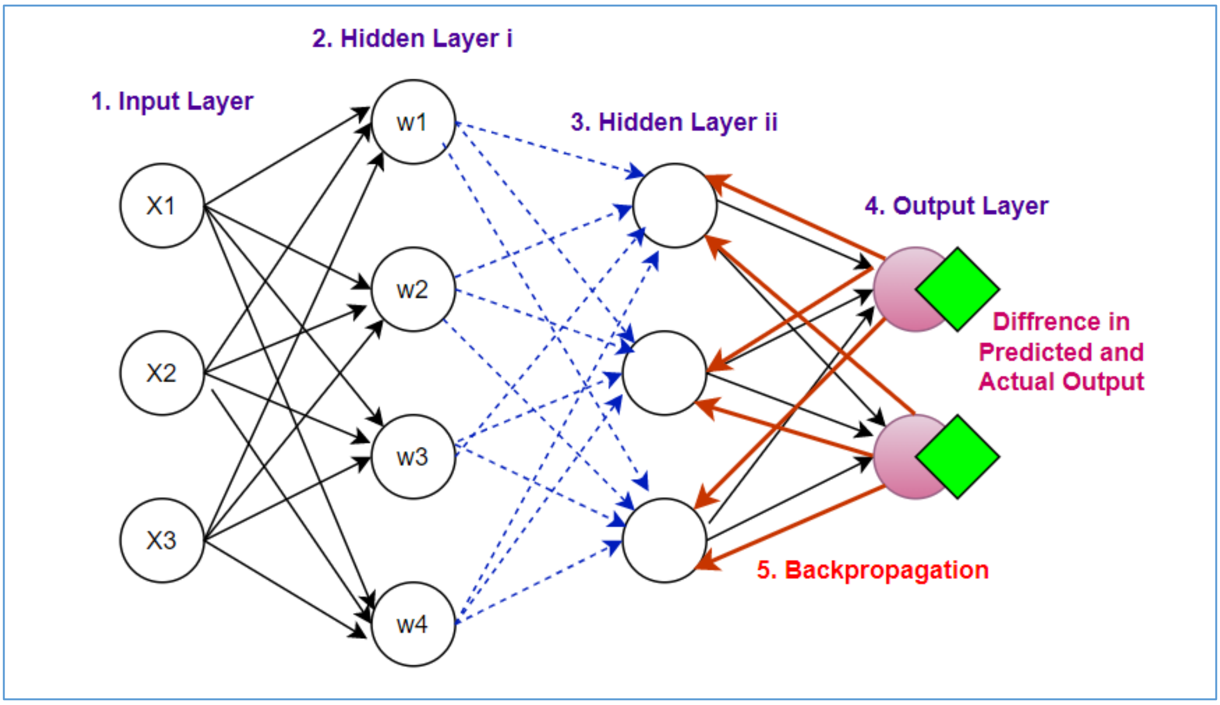


Figure 3.13 - Backpropagation

7. Application of Loss Function:

To make model performance better loss function is use. Loss function calculate difference between predicted output and actual output in real number value. The precision of the model is higher when lower value is having for loss value.

Error = Actual Output - Desired Output

Loss values can calculate in three ways such as Gradient Decent, Stochastic Decent and Mini-Batch Stochastic Descent. In here we are considering gradient decent only as taking whole set of value at once. In gradient descent we calculate slope of the point getting down to minimum cost/loss value. When slope is 0, cost value is minimum. Slope point upward is positive and downward is negative. To get the right learning rate use to get minimum value. Learning rate algorithms reduce cost function. Figure 3.14 shows how gradient decent works.

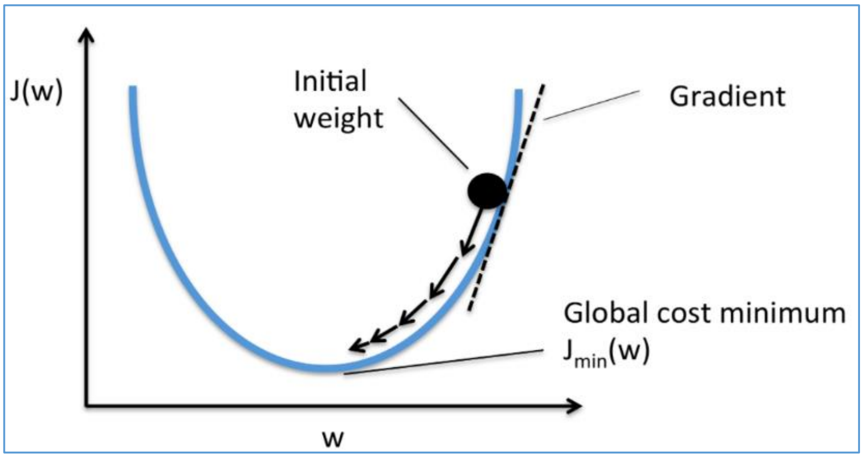


Figure 3.14 – Gradient Decent (Source - Nellur, 2020)

3.4.2 Support Vector Machine

“Support Vector Machine” (SVM) is a supervised machine learning algorithm. SVM can be used for classification and regression problems. SVM plots each data item as a point in n-dimensional space. Here n is the number of features. Value of each feature belongs to a particular coordinate. Classification was performed by finding hyperplanes that divide two classes (Ray, n.d.).

Hyperplane

Hyperplane is the best decision boundary. Features penetrated in the data set decide the dimension of the hyperplane. Hyperplane will be a straight line if we have two features. If we have more features, the hyperplane will be a 2D plane. Maximum distance between data points calculate maximum margin. Hyperplane created by considering maximum margin.

Support Vector

SVM selects extreme points known as support vectors. The extreme points are the points closest to the hyperplane. These vectors support hyperplanes and hence we call them support vectors.

Figure 3.15 shows components of linear SVM.

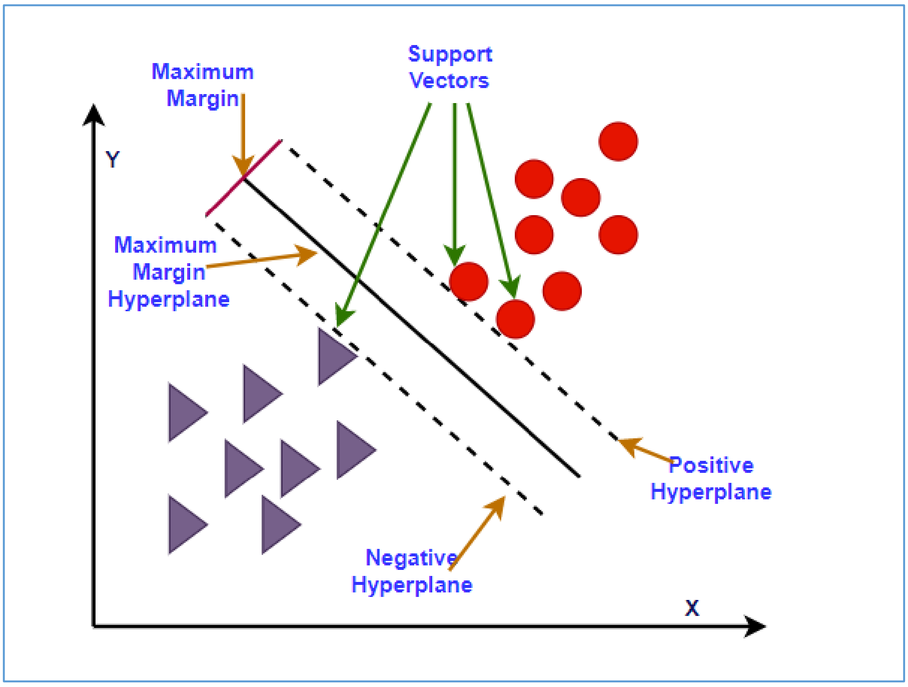


Figure 3.15 - SVM decision boundary or hyperplane

SVM are categorized into two types as linear and nonlinear. Using a straight line, we can divide data into two dimensions in linear SVM. We cannot separate if the data are arranged in non-linearly. Linear data used two dimensions X and Y. For nonlinear data Z is added additionally. We can use both linear and nonlinear methods and select the more accurate model. Figure 3.16 shows an example for Nonlinear.

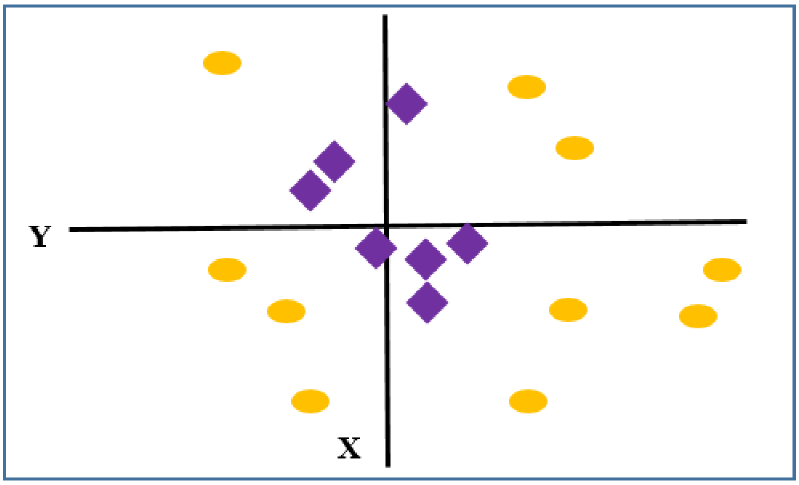


Figure 3.16 - Nonlinear – SVM

3.4.3 Model Validation

Confusion Matrix - Confusion Matrix is widely used performance measurement in classification problems. Figure 3.17 shown basic component of a confusion matrix.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Figure 3.17 – Confusion Matrix

- True Positive – Predicted value is positive and equal to actual positive value
- True Negative - Predicted value is negative and equal to actual negative value
- False Positive (Type 1 Error) – Predicted as positive and actual value is not positive
- False Negative (Type 2 Error) - Predicted as negative and actual value is not negative.
- Recall – Considering positive classes how many correct predictors.

  Recall = TP /(TP+FN) (2)

- Accuracy - Considering all classes (positive and negative), how many correct predictors.

  Accuracy = (TP+TN) / Total (3)

- Precision – Considering all positive predicted classes how many correct positive actuals.

  Precision = TP/(TP+FP) (4)

- Recall, Accuracy, precision values should be high for better performance.
- F1 Measure – It is difficult to compare precision and recall when they are in high and low or vice versa. Precision and recall can be measure at same time by using F-score.

  F1 Measure = (2 * Recall * Precision) / (Recall + Precision) (5)

Mean Absolute Error (MAE) – the gap between the measured and "true" values

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \qquad (6)$$

CHAPTER 4

IMPLEMENTATION AND RESULTS

This chapter describes implementation and results evaluation activities in detail. List of activities discussed in here are explanatory analysis of data, data preparation activities, models building, evaluation of models and deployments.

4.1 Explanatory Data Analysis

Graphical and numerical representation of data provide better insight about particular data set. Graphical representation of our data set is described below.

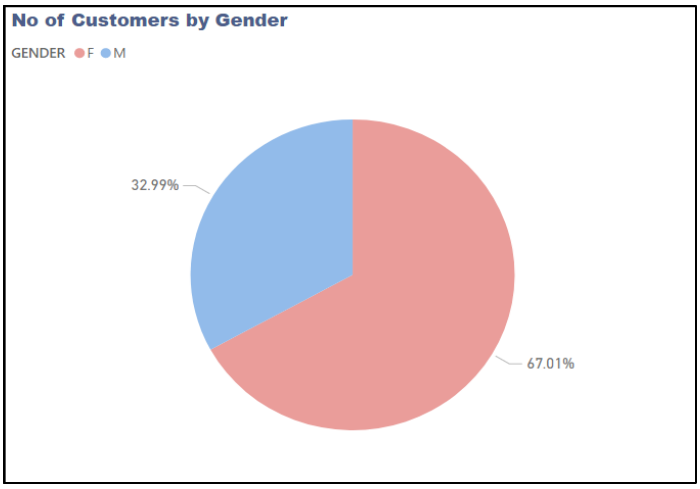


Figure 4.1– No of Customers by Gender

According to the figure 4.1 distribution of gender of the customers are 67 % female and 32% are male.

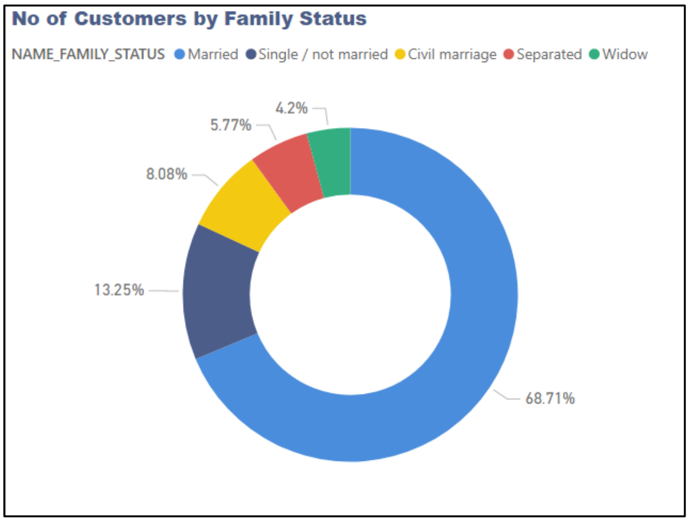


Figure 4.2– No of Customers by Family Status

According to the 4.2 graph No of Customers by family status 69% are married, 13% are single, 8% are civil marriage, 5% are separated and 4% are widows.

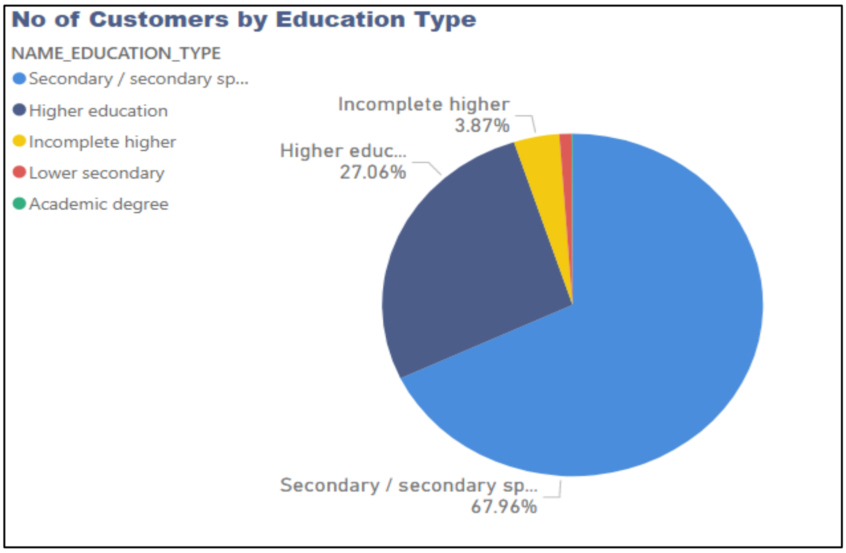


Figure 4.3– No of Customers by Education Type

As shown in figure 4.3 No of Customers by Education Type figure 68% have secondary education, 27% are have higher education and 4% are have incomplete higher education.

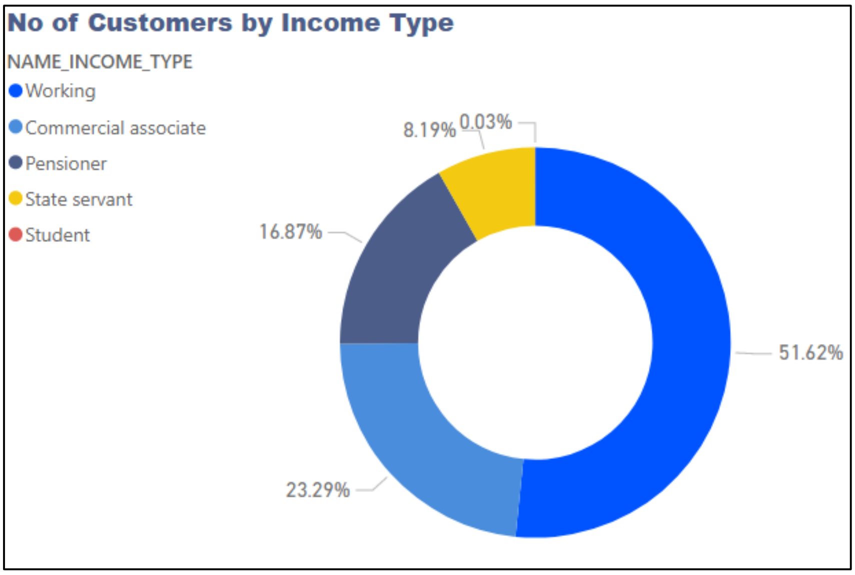


Figure 4.4 – No of Customers by Income Type

According to the graph 4.4 No of Customers by Income Type, 62% are working, 23% are commercial associates, 16% are pensioner, 8% State servant.

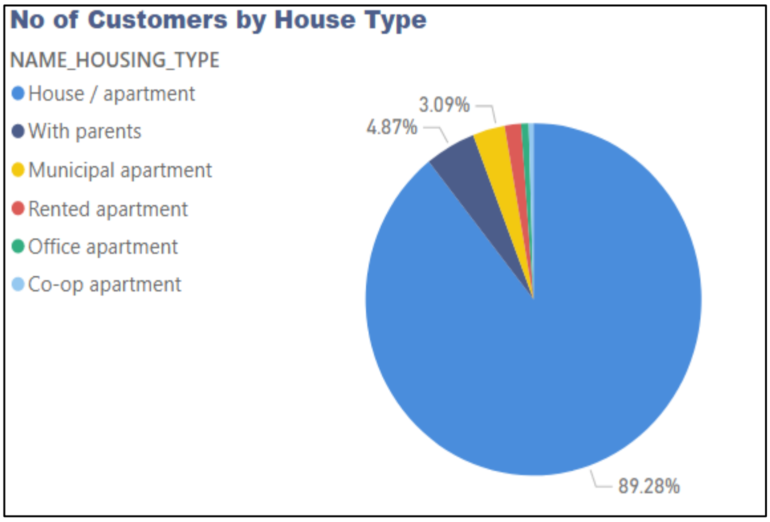


Figure 4.5– No of Customers by House Type

According to the graph 4.5 No of Customers by House Type, 89% are have house, 5% living with parents, 3% are live in municipal apartment.

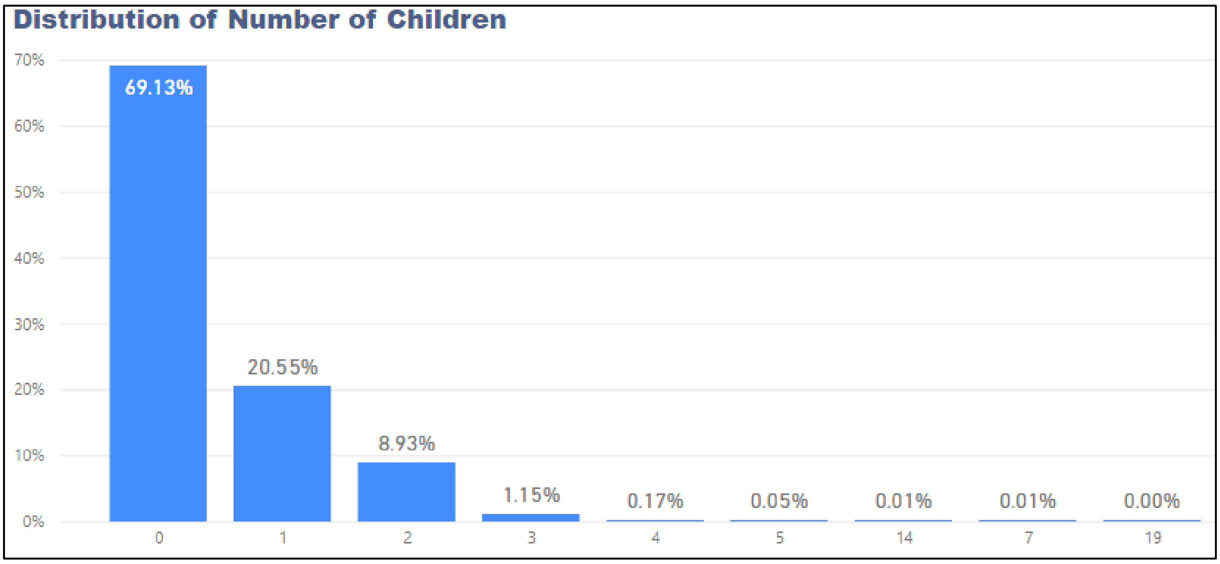


Figure 4.6– Distribution of Number of Children

As shown in 4.6 graph Distribution of Numbers of Children, 69 % are not have a child. 20% have one child, 9% have two children and 1% have 3 children. Remaining 0.5% have more than 3 children.

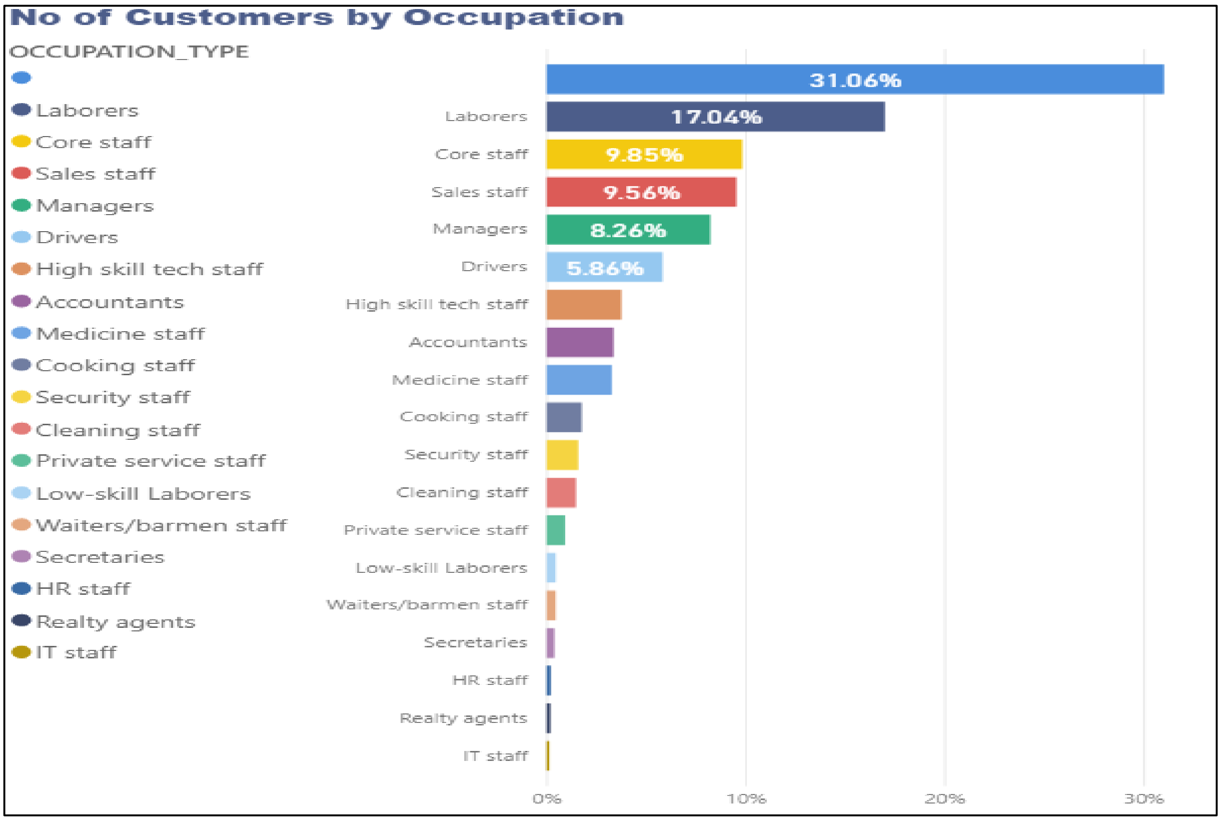


Figure 4.7– No of Customers by Occupation Type

According to the above graph 4.7 Distribution of occupation, 31% have missing values (occupation not mentioned), 17 % are labors, 10% are core staff, 10% sales staff, 8% are managers and 6% are high skill tech staff.

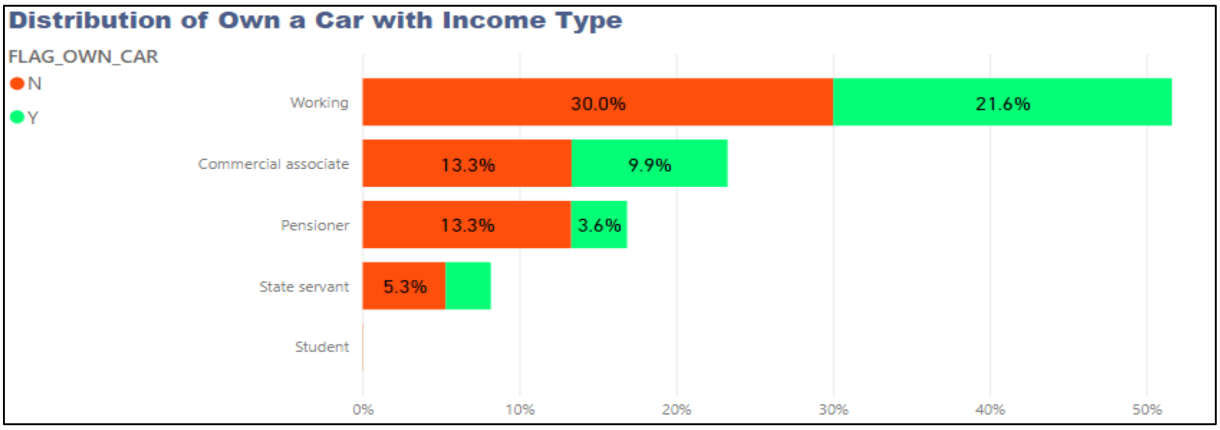


Figure 4.8– Distribution of Own a Car with Income Type

According to the above graph 4.8 Distribution of own a car with income type, 50 % are working customers and 30 % out of them does not own a car and 20% own a car. 20 % are commercial associates and 13 % out of them does not own a car and 10% own a car. 15 % are pensioner and 13 % out of them does not own a car and 4% own a car.

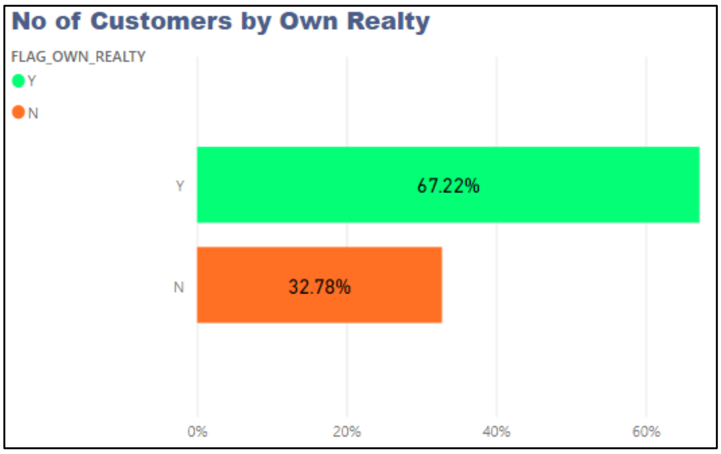


Figure 4.9– No of Customers by Own Realty

As shown in 4.9 graph No of Customers by Own Realty, 67% are own a realty and 33% are not own a realty.

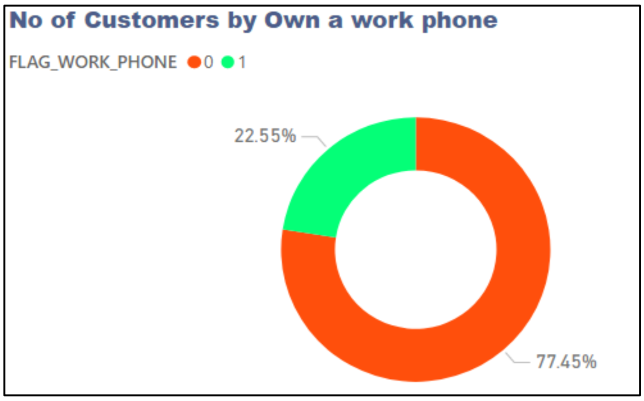


Figure 4.10– No of Customers by own a work phone

According to the graph 4.10 No of Customers by Own a work phone, 77% are not own a work phone and 23% are own a work phone.

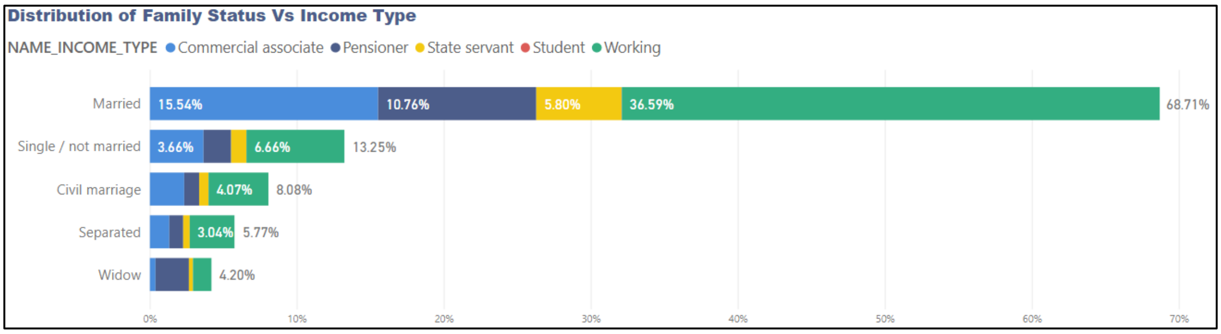


Figure 4.11– Distribution of Family Status Vs Income Type

As shown in the graph 4.11 68% are married and 37% of them are working, 15% are commercial associates, 11% are pensioner and 6% are state servants.

4.2 Data Preparation Activities

Data preparation activities related to our project discuss separately below.

4.2.1 Clean Data

ID column contains white spaces. Joining two data set with white spaces does not give correct aggregation. Removed white spaces from ID column in both data set. ID column converted as string column.

DAYS_EMPLOYED column count backwards from present day (0). Values contains negative and positive both. Positive mean the person currently unemployed. Positive data values are set to 0 and convert negative values to positive value by multiplying -1 to bring into standard format.

4.2.2 Handling Missing Value

In this data set there are missing values in OCCUPATION_TYPE column and hence this is a categorical column replaced missing values with ‘Other’.

4.2.3 Construct Data

This data set doesn’t contain direct class label. Derived a variable from applicant information data set as customer is good or bad by using credit card payment history details.

Methodology of Deriving a Dependent Variable (Class Variable)

By using customer’s credit history data dependent variable was generated. There are 8 different payment status and they are listed below.

- C: paid off that month
- X: No loan for the month
- 0: 1-29 days past due
- 1: 30-59 days past due
- 2: 60-89 days overdue
- 3: 90-119 days overdue
- 4: 120-149 days overdue
- 5: Overdue or bad debts, write-offs for more than 150 days

After 60 days or more customers’ delinquent status become overdue and most companies involve in taking legal actions to start debt collection. Hence, the overdue status codes 3, 4, and 5 consider as bad customer status. As shown in table 4.1, customers who are paid off that month, those who have not taken loan for the month, those who are not paid and past due for 1 to 29 days and 30-59 days past due consider as good customers. Payments are past due from 60th day and above consider as defaulters to minimize class imbalance problem. After selecting default codes new column “Final_Label” created as 0 is for a good customer and 1 is for bad customer considering below mentioned logic.

Table 4.1 - Class Label Classification Table

Data Value	Data Value Description	Label
X	C: paid off that month	0
C	X: No loan for the month	0
0	0: 1-29 days past due	0
1	1: 30-59 days past due	0
2	2: 60-89 days overdue	1
3	3: 90-119 days overdue	1
4	4: 120-149 days overdue	1
5	5: Overdue or bad debts, write-offs for more than 150 days	1

Aggregate table generated with 4 columns namely ID, NO_OF_GOOD, NO_OF_BAD, FINAL_LABEL. Unique IDs taken from credit record data set and counted no of bad and good status separately. Considering NO_OF_BAD column another column FINAL_LABEL created for below logic. Table 4.2 shows sample of aggregate table.

If 'NO_OF_BAD' < 1 then 1

If 'NO_OF_BAD' > 0 then 0

Table 4.2 - Sample Aggregated Table

ID	NO_OF_GOOD	NO_OF_BAD	FINAL_LABEL
5001711	4	0	0
5001712	19	0	0
5001713	22	0	0
5001718	37	2	1
5001715	60	0	0
5001720	29	7	1

Further derived two new columns from DAYS_BIRTH and DAYS_EMPLOYED as AGE_IN_YEARS and EMPLOYED_IN_YEARS by dividing 365 (making to years) because of days are make less sense in business domain.

4.2.4 Integrated Data

In here we have two data set combined together. Below figure 4.12 illustrated combining of two data set.

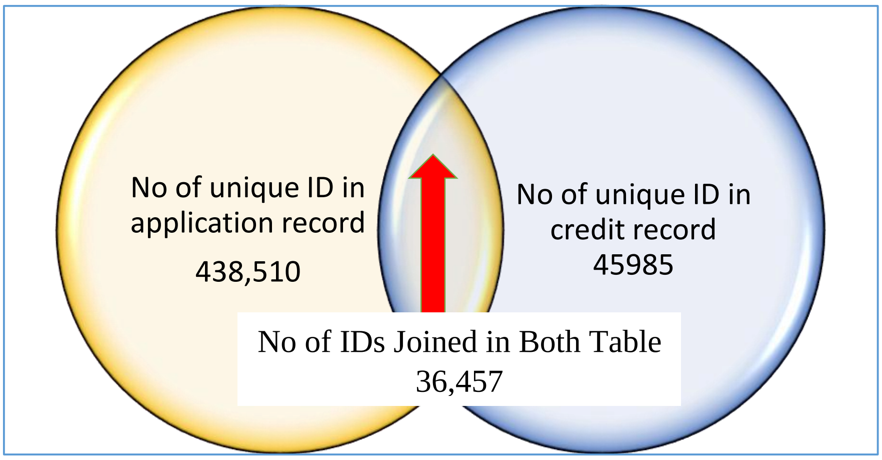

No of unique ID in application record
438,510
No of unique ID in credit record
45985
No of IDs Joined in Both Table
36,457

Figure 4.12- Selected Data Set

Credit record data sets and application record data sets are merged together to generate final data set. After identification of class label there are 615 bad customers and 35841 good customers. Final data set consider for model building is 35841.

4.2.5 Outlier Removals

To identify outlier boxplot and histogram of continuous data column separately analysis. Following figures (4.13, 4.14, 4.15, 4.16) shows the boxplot and relevant histogram for each value of outlier analysis.

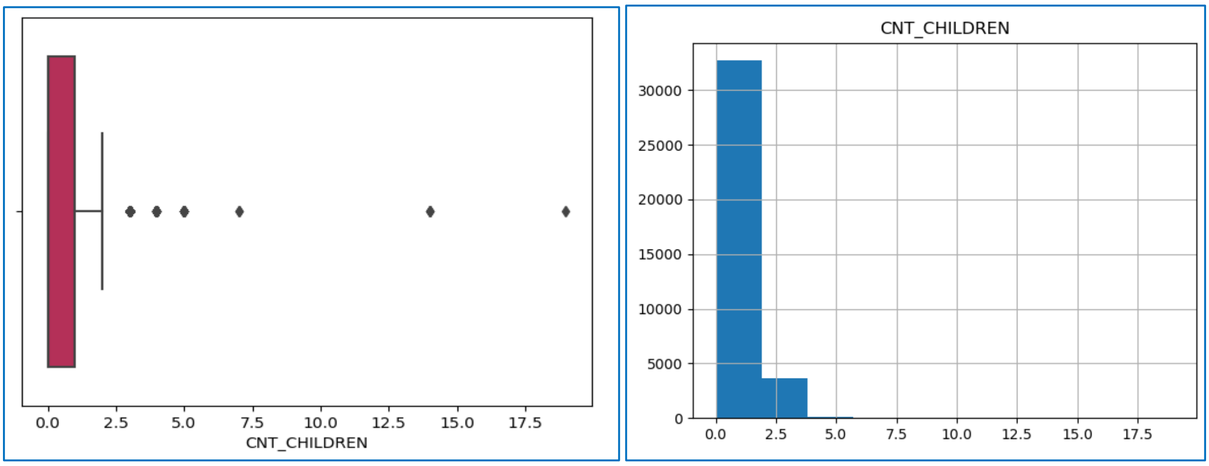


Figure 4.13 - Boxplot and Histogram for No of Children

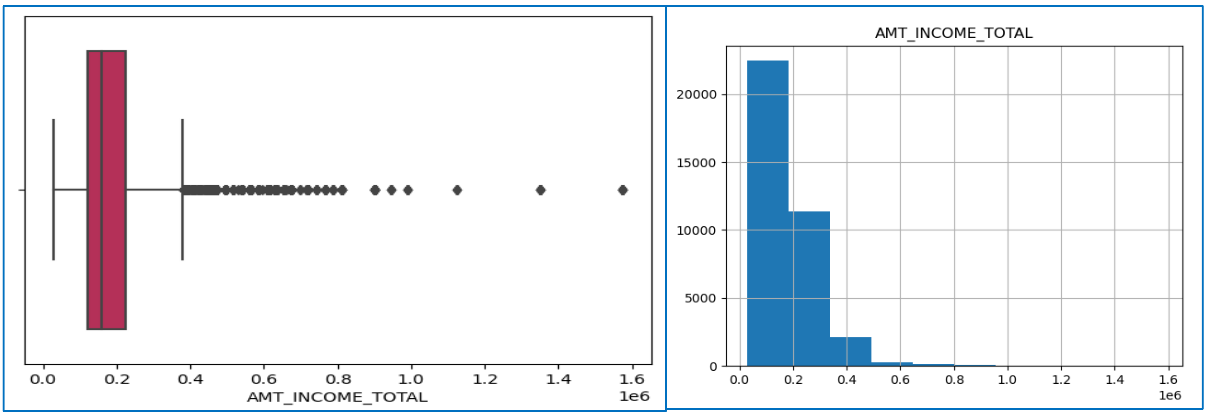


Figure 4.14 - Boxplot and Histogram for Income Total

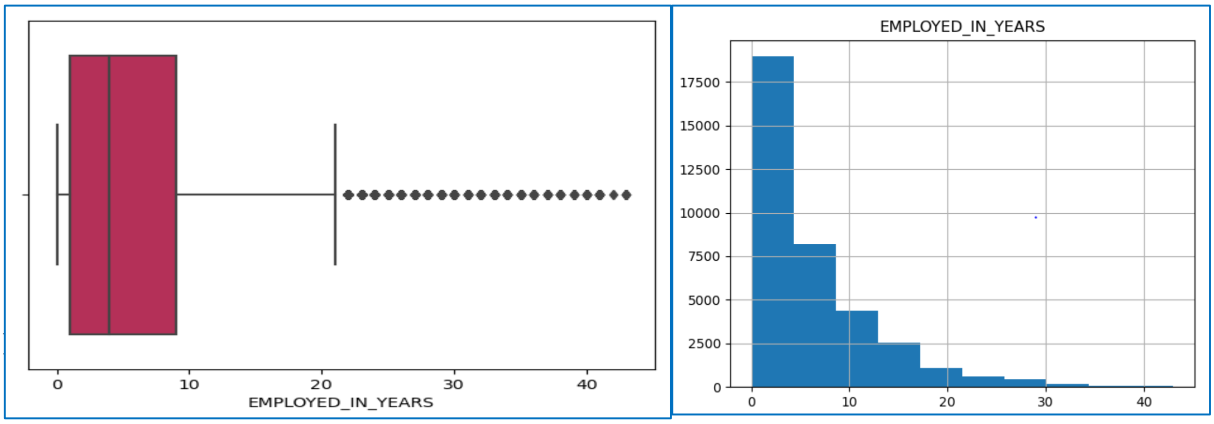


Figure 4.15 - Boxplot and Histogram for Employed in Years

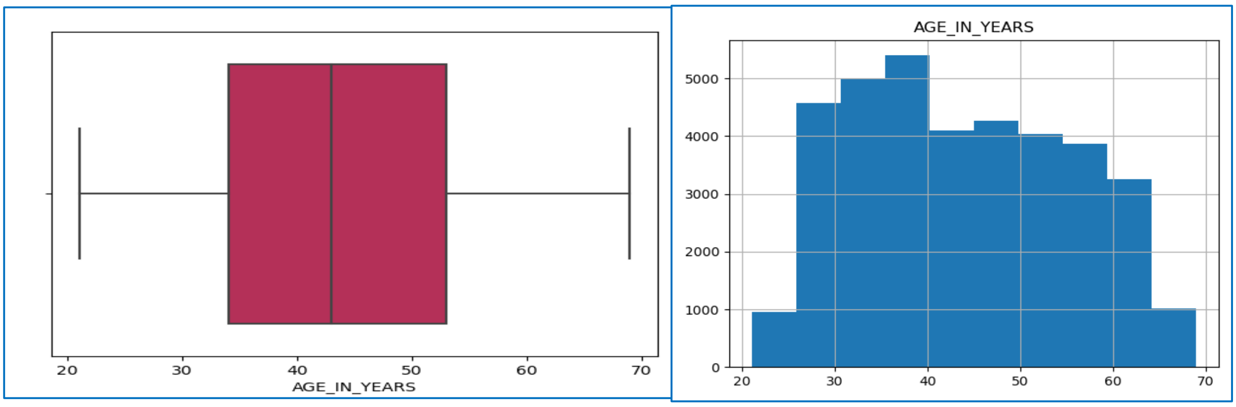


Figure 4.16 - Boxplot and Histogram for Age in Years

Outliers in AMT_INCOME_TOTAL, CNT_CHILDREN, AGE_IN_YEARS and EMPLOYED_IN_YEARS were deleted by using statistical method Inter Quartile Range (IQR) and removed outliers from the data set. Final data set is 33,140.

4.2.6 Encoding Categorical Data

There are 3 methods used to convert categorical features to numerical as shown in figure 4.17. In ordinal encoding the ordinal nature of the variable is considered. Sequence of integers will assign. Encoding with get_dummies maps category columns to a vector containing 0 and 1 denoting the presence or absence of the feature. Number of vectors is dependent on the number of categories in the column value. Under binary encoding map category column into 2 new columns as 0 and 1 denoting the presence and absence of the feature.

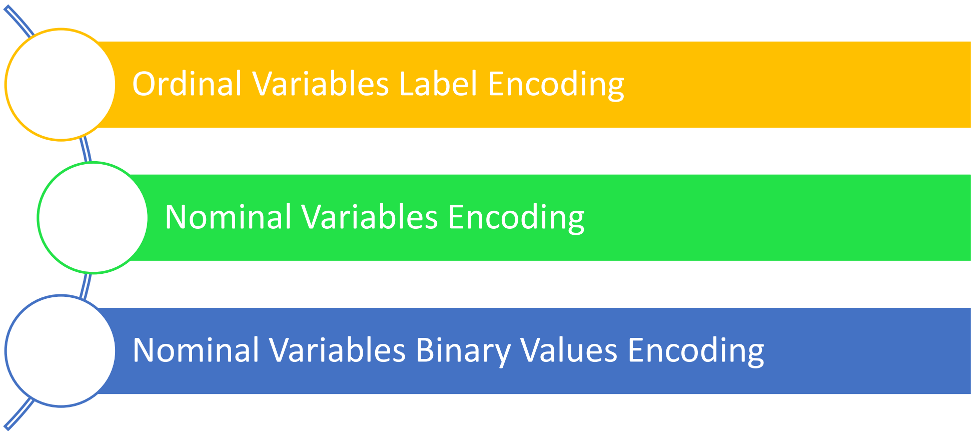

Ordinal Variables Label Encoding
Nominal Variables Encoding
Nominal Variables Binary Values Encoding

Figure 4.17 - Encoding Types

Each features treated differently as shown in the below table 4.3.

Table 4.3 - Categorical Variables with encoding methods

Feature name	Data Type	Encoding Methods
CODE_GENDER	Categorical – nominal (binary)	Python package category_encoders BinaryEncoder
FLAG_OWN_CAR	Categorical – nominal (binary)	Python package category_encoders BinaryEncoder
FLAG_OWN_REALTY	Categorical – nominal (binary)	Python package category_encoders BinaryEncoder
NAME_INCOME_TYPE	Categorical – nominal	Python package pandas get_dummies
NAME_EDUCATION_TYPE	Categorical - ordinal	Python package category_encoders OrdinalEncoder
NAME_FAMILY_STATUS	Categorical - nominal	Python package pandas get_dummies
NAME_HOUSING_TYPE	Categorical - nominal	Python package pandas get_dummies
OCCUPATION_TYPE	Categorical - nominal	Python package pandas get_dummies

Variable list with data type after application of categorical encoding as follows.

```
FINAL_LABEL                    int32
CODE_GENDER_0                  int64
CODE_GENDER_1                  int64
FLAG_OWN_CAR_0                 int64
FLAG_OWN_CAR_1                 int64
FLAG_OWN_REALTY_0              int64
```

```
FLAG_OWN_REALTY_1                                int64
NAME_EDUCATION_TYPE                              int32
FLAG_WORK_PHONE_0                                int64
FLAG_WORK_PHONE_1                                int64
FLAG_PHONE_0                                     int64
FLAG_PHONE_1                                     int64
FLAG_EMAIL_0                                     int64
FLAG_EMAIL_1                                     int64
NAME_INCOME_TYPE_Commercial associate            int32
NAME_INCOME_TYPE_Pensioner                       int32
NAME_INCOME_TYPE_State servant                   int32
NAME_INCOME_TYPE_Student                         int32
NAME_INCOME_TYPE_Working                         int32
NAME_FAMILY_STATUS_Civil marriage                int32
NAME_FAMILY_STATUS_Married                       int32
NAME_FAMILY_STATUS_Separated                     int32
NAME_FAMILY_STATUS_Single / not married          int32
NAME_FAMILY_STATUS_Widow                         int32
NAME_HOUSING_TYPE_Co-op apartment                int32
NAME_HOUSING_TYPE_House / apartment              int32
NAME_HOUSING_TYPE_Municipal apartment            int32
NAME_HOUSING_TYPE_Office apartment               int32
NAME_HOUSING_TYPE_Rented apartment               int32
NAME_HOUSING_TYPE_With parents                   int32
OCCUPATION_TYPE_Accountants                      int32
OCCUPATION_TYPE_Cleaning staff                   int32
OCCUPATION_TYPE_Cooking staff                    int32
OCCUPATION_TYPE_Core staff                       int32
OCCUPATION_TYPE_Drivers                          int32
OCCUPATION_TYPE_HR staff                         int32
OCCUPATION_TYPE_High skill tech staff            int32
OCCUPATION_TYPE_IT staff                         int32
OCCUPATION_TYPE_Laborers                         int32
OCCUPATION_TYPE_Low-skill Laborers               int32
OCCUPATION_TYPE_Managers                         int32
OCCUPATION_TYPE_Medicine staff                   int32
OCCUPATION_TYPE_Other                            int32
OCCUPATION_TYPE_Private service staff            int32
OCCUPATION_TYPE_Realty agents                    int32
OCCUPATION_TYPE_Sales staff                      int32
OCCUPATION_TYPE_Secretaries                      int32
OCCUPATION_TYPE_Security staff                   int32
OCCUPATION_TYPE_Waiters/barmen staff             int32
```

4.2.7 Feature Selection

Before application of filter base statistical method, distribution of numerical variables was analyzed. Figure 4.18 illustrated distribution of single variable and relationship between two variables.

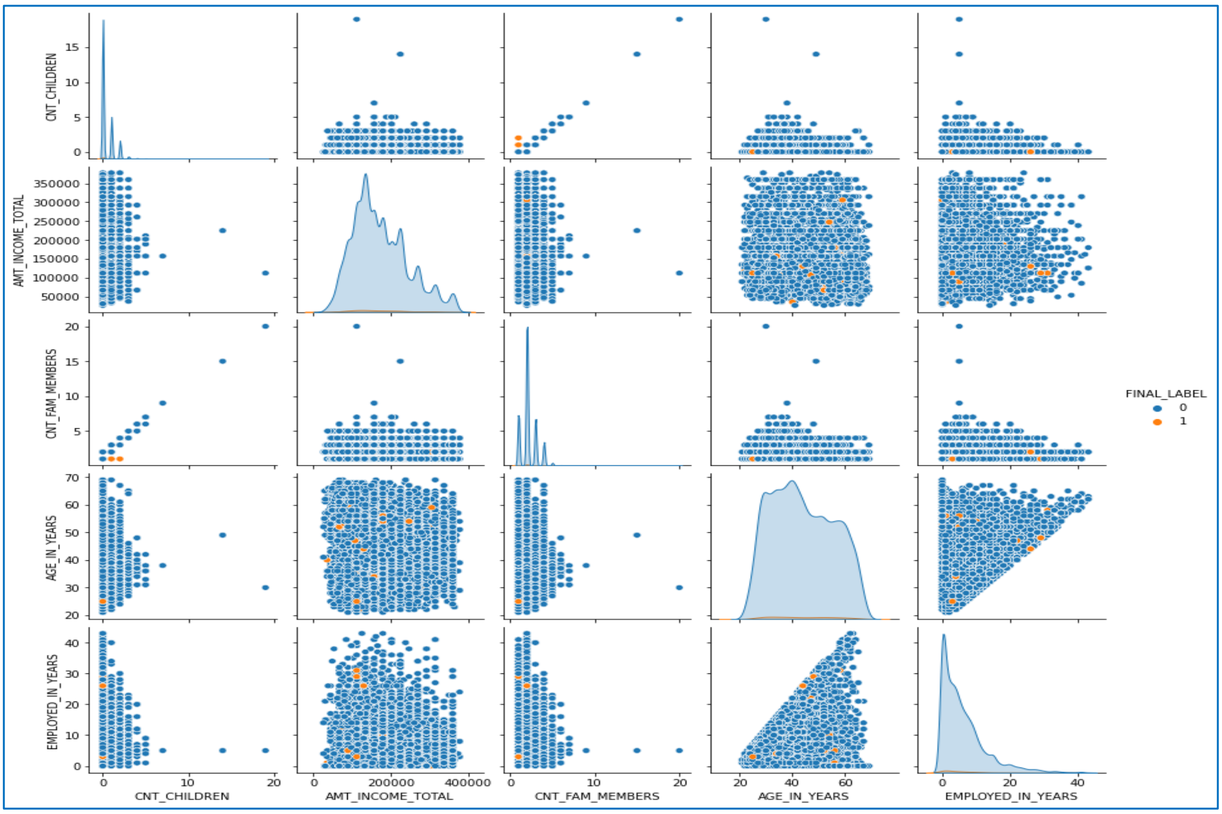


Figure 4.18 - Numerical Variable Pair plot

Correlation Based Feature Selection

Pearson Correlation Coefficient – To measure relationship between linearly related variables Pearson r correlation is the most widely used. Figure 4.19 and 4.20 shown correlation in numerical data and they have weak relationship.

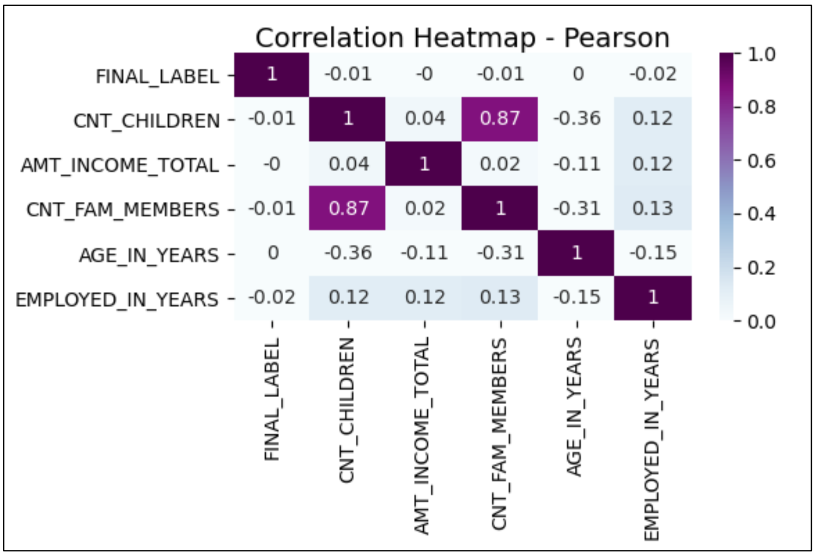


Figure 4.19 - Correlation heat map for Pearson

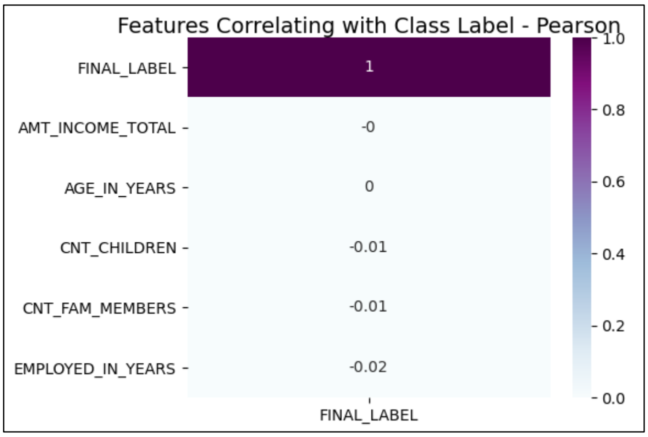


Figure 4.20 - Pearson correlation values with Class Label

Spearman Rank Correlation - To measure the degree of association between two variables Spearman rank correlation is use. Spearman rank correlation is a non-parametric test. Figure 4.21 and 4.22 shown spearman correlation in numerical data.

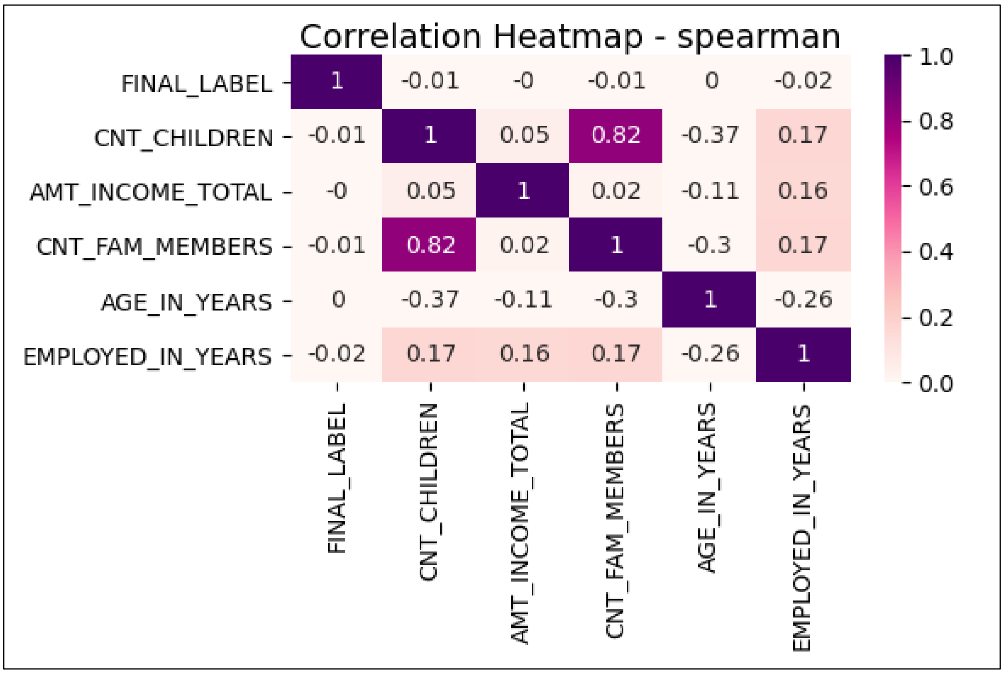


Figure 4.21- Correlation heat map for Spearman

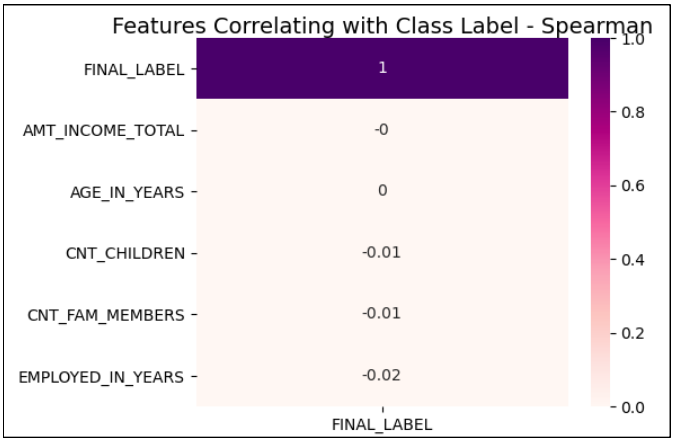


Figure 4.22 - Spearman correlation values with Class Label

Information Gain

sklearn.feature_selection import mutual_info_classif use to calculate information gain and figure 4.23 shown categorical features.

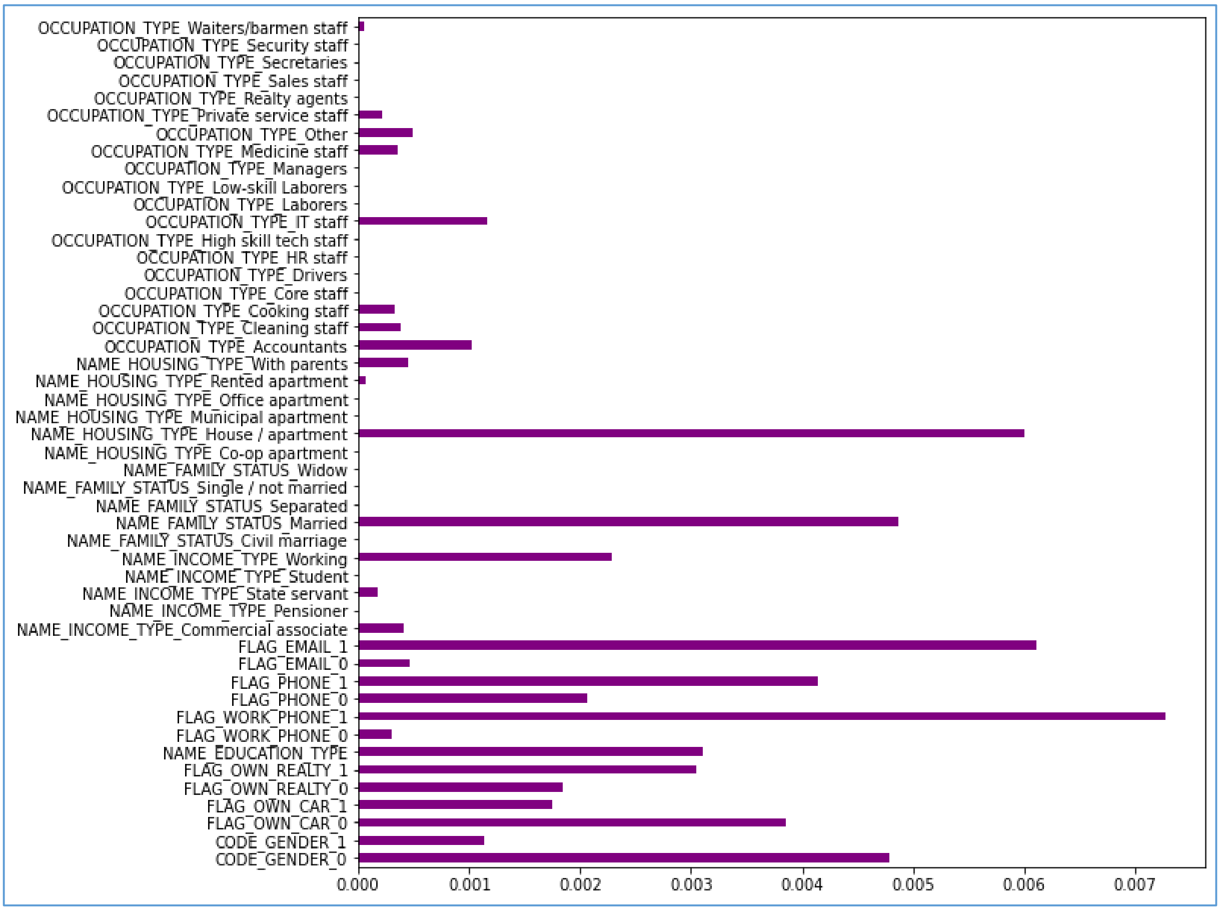


Figure 4.23 - Categorical features with Mutual Information gain

Correlation > 0.1 and mutual information gain > 0.001 features were selected as input to the models. Finalized feature set of categorical variables are shown in Table 4.4.

Table 4.4 - Finalized Categorical Viable List

#	Feature Name	Value
1	FLAG_WORK_PHONE_1	0.007271
2	FLAG_EMAIL_1	0.006114
3	NAME_HOUSING_TYPE_House apartment	0.005997
4	NAME_FAMILY_STATUS_Married	0.004869
5	CODE_GENDER_0	0.004784
6	FLAG_PHONE_1	0.004142
7	FLAG_OWN_CAR_0	0.003852
8	NAME_EDUCATION_TYPE	0.003109
9	FLAG_OWN_REALTY_1	0.003042
10	NAME_INCOME_TYPE_Working	0.002278
11	FLAG_PHONE_0	0.002056
12	FLAG_OWN_REALTY_0	0.001848
13	FLAG_OWN_CAR_1	0.001744
14	OCCUPATION_TYPE_IT staff	0.001161
15	CODE_GENDER_1	0.001133
16	OCCUPATION_TYPE_Accountants	0.001018

Final Feature set from numerical and categorical are listed below.

1. CNT_CHILDREN
2. AMT_INCOME_TOTAL
3. CNT_FAM_MEMBERS
4. AGE_IN_YEARS
5. EMPLOYED_IN_YEARS
6. FLAG_WORK_PHONE_1
7. FLAG_EMAIL_1
8. NAME_HOUSING_TYPE_House apartment
9. NAME_FAMILY_STATUS_Married
10. CODE_GENDER_0
11. FLAG_PHONE_1
12. FLAG_OWN_CAR_0
13. NAME_EDUCATION_TYPE
14. FLAG_OWN_REALTY_1
15. NAME_INCOME_TYPE_Working
16. FLAG_PHONE_0
17. FLAG_OWN_REALTY_0
18. FLAG_OWN_CAR_1
19. OCCUPATION_TYPE_IT staff
20. CODE_GENDER_1
21. OCCUPATION_TYPE_Accountants

4.2.8 Feature Scaling

We have applied Standard scaler for numerical features. Below numerical features have been transformed.

- AMT_INCOME_TOTAL
- NAME_EDUCATION_TYPE
- CNT_FAM_MEMBERS
- EMPLOYED_IN_YEARS
- AGE_IN_YEARS
- CNT_CHILDREN

4.2.9 Handling Imbalance Data

We can see class imbalance problem in here. After data wangling process we have 561 bad customers and 32,580 good customers. In our data set Bad customers are significantly lower than good customer as shown in figure 4.24.

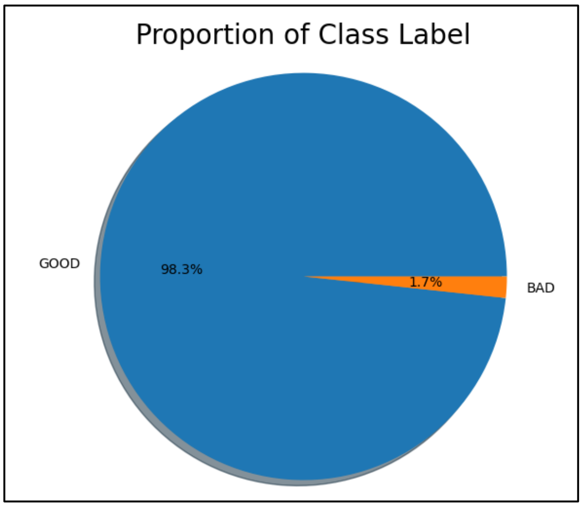


Figure 4.24 - Proportion of Class label

In here we have use synthetic sampling approach, SMOTE (Synthetic Minority Oversampling Technique) to handling imbalanced data. Our data set have 32,586 records and dimensions are not in very high.

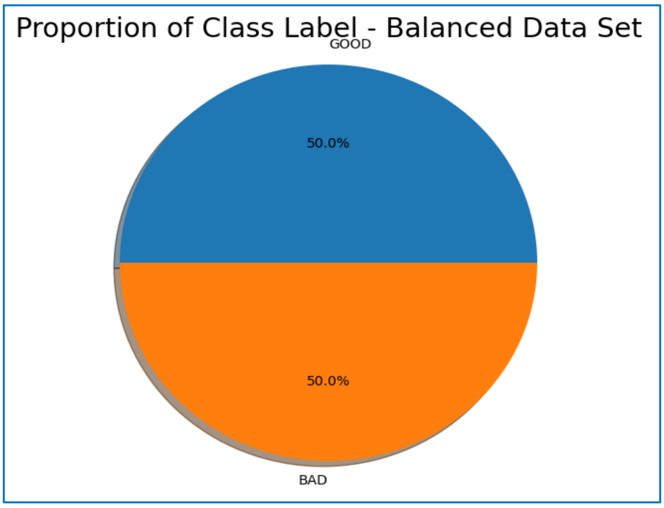


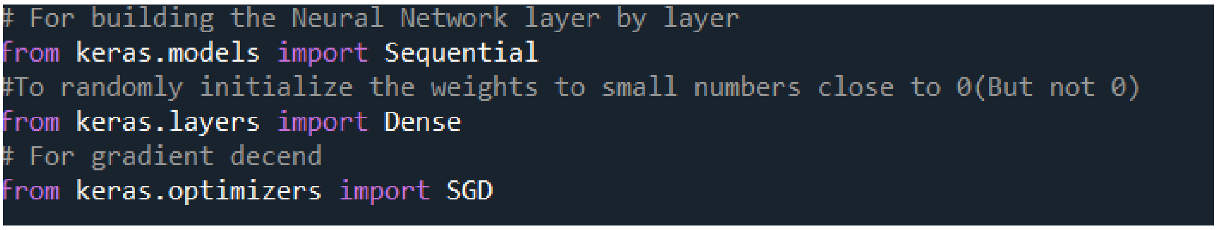
Figure 4.25 - Data set after application of SMOTE

Figure 4.25 shows proportionate of class variable after application of SMOTE. To carry out this task imbalanced-learn Python library use.

4.3 Model Building

4.3.1 Application of Artificial Neural Network

We use Keras library to develop our ANN model as shown in the figure 4.26. In keras model build as a sequence of layers by adding layers one at a time.



```
# For building the Neural Network layer by layer
from keras.models import Sequential
#To randomly initialize the weights to small numbers close to 0(But not 0)
from keras.layers import Dense
# For gradient decend
from keras.optimizers import SGD
```

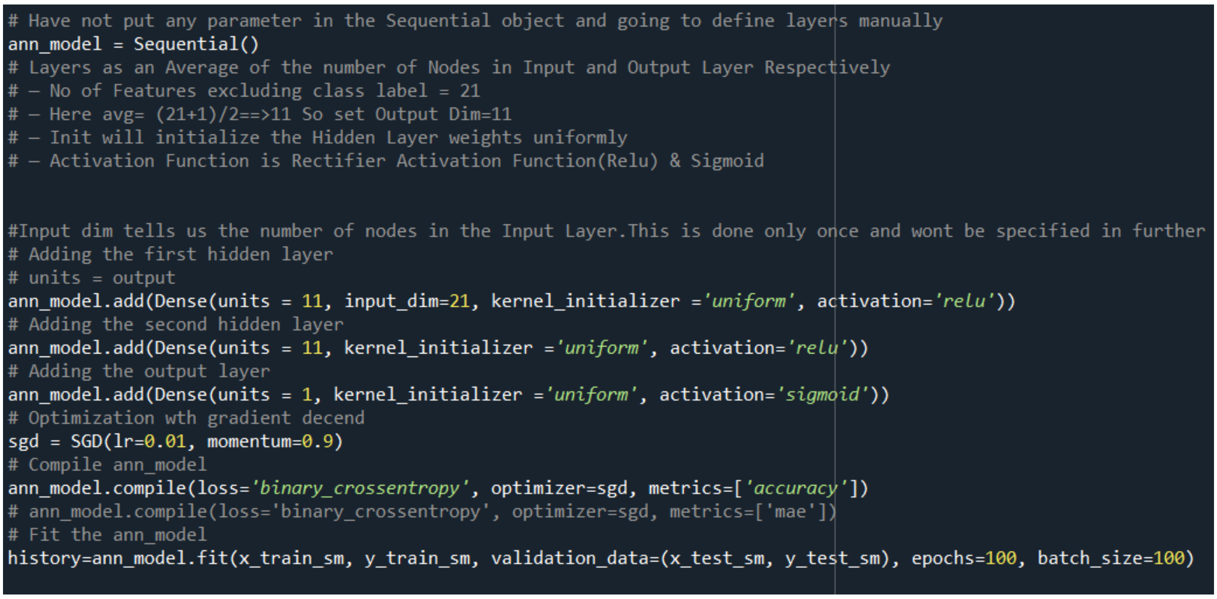
Figure 4.26 – Python Libraries

Main parameters of Keras model building as follows.

- Units = Number of variables
- Number of Nodes in the input layer = 21 and activation function = relu.
- Number of Nodes in the second layer = 11 nodes and activation function = relu
- Number of Nodes in the output layer = 1 activate function = sigmoid
- The batch size = Number of samples processed before the model is updated
    - Batch size must be greater ≥ 1 or ≤ Number of samples in the data set

- Number of Epochs = Number of complete passes through the training dataset
- Weights initialization = Uniform distribution used. (kernel initializer ='uniform')
- Activate function = Rectifier Activation Function (Relu) & Sigmoid Model fitting parameters
- Loss = binary_crossentropy
- Optimizer = Stochastic Gradient Descent
- Metrics = accuracy

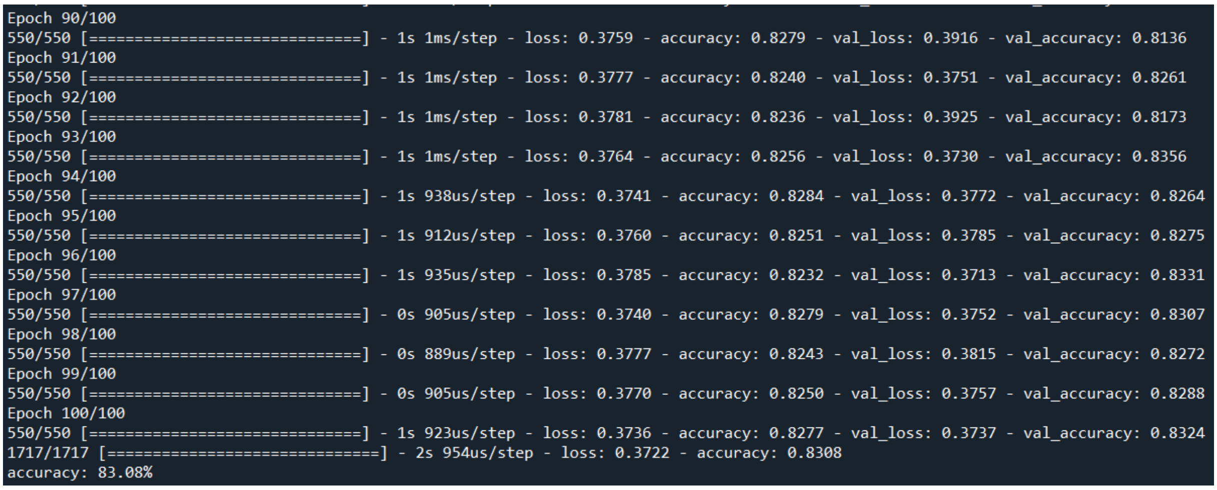
Our training data set contain 51,510 sample. We have chosen batch size as 100 and epoch as 100. In here data set will be divided into 515.1 batches each with 100 samples. After each batch of 515 samples, the model weight will be updated. In the other words one epoch involved with 515 batches. In entire training process the model will pass through entire data set 100 times which is total of 51,510 batches. Figure 4.27 shows parameters of model building



```
# Have not put any parameter in the Sequential object and going to define layers manually
ann_model = Sequential()
# Layers as an Average of the number of Nodes in Input and Output Layer Respectively
# — No of Features excluding class label = 21
# — Here avg= (21+1)/2==>11 So set Output Dim=11
# — Init will initialize the Hidden Layer weights uniformly
# — Activation Function is Rectifier Activation Function(Relu) & Sigmoid


#Input dim tells us the number of nodes in the Input Layer.This is done only once and wont be specified in further
# Adding the first hidden layer
# units = output
ann_model.add(Dense(units = 11, input_dim=21, kernel_initializer ='uniform', activation='relu'))
# Adding the second hidden layer
ann_model.add(Dense(units = 11, kernel_initializer ='uniform', activation='relu'))
# Adding the output layer
ann_model.add(Dense(units = 1, kernel_initializer ='uniform', activation='sigmoid'))
# Optimization wth gradient decend
sgd = SGD(lr=0.01, momentum=0.9)
# Compile ann_model
ann_model.compile(loss='binary_crossentropy', optimizer=sgd, metrics=['accuracy'])
# ann_model.compile(loss='binary_crossentropy', optimizer=sgd, metrics=['mae'])
# Fit the ann_model
history=ann_model.fit(x_train_sm, y_train_sm, validation_data=(x_test_sm, y_test_sm), epochs=100, batch_size=100)
```

Figure 4.27 - Parameters of the ANN model

Preferably, we would like to have accuracy to be 1.0 (100%) and loss to be zero in our ML model. However, most machine learning solution does not always give 100 percent accuracy. Thus the goal is to achieve highest accuracy and lowest loss in our data set. Figure 4.28 shows last 10 records of model execution. Each of 100 epoch printing accuracy and loss of training data set as well as testing (validation) data set.



```
Epoch 90/100
550/550 [==============================] - 1s 1ms/step - loss: 0.3759 - accuracy: 0.8279 - val_loss: 0.3916 - val_accuracy: 0.8136
Epoch 91/100
550/550 [==============================] - 1s 1ms/step - loss: 0.3777 - accuracy: 0.8240 - val_loss: 0.3751 - val_accuracy: 0.8261
Epoch 92/100
550/550 [==============================] - 1s 1ms/step - loss: 0.3781 - accuracy: 0.8236 - val_loss: 0.3925 - val_accuracy: 0.8173
Epoch 93/100
550/550 [==============================] - 1s 1ms/step - loss: 0.3764 - accuracy: 0.8256 - val_loss: 0.3730 - val_accuracy: 0.8356
Epoch 94/100
550/550 [==============================] - 1s 938us/step - loss: 0.3741 - accuracy: 0.8284 - val_loss: 0.3772 - val_accuracy: 0.8264
Epoch 95/100
550/550 [==============================] - 1s 912us/step - loss: 0.3760 - accuracy: 0.8251 - val_loss: 0.3785 - val_accuracy: 0.8275
Epoch 96/100
550/550 [==============================] - 1s 935us/step - loss: 0.3785 - accuracy: 0.8232 - val_loss: 0.3713 - val_accuracy: 0.8331
Epoch 97/100
550/550 [==============================] - 0s 905us/step - loss: 0.3740 - accuracy: 0.8279 - val_loss: 0.3752 - val_accuracy: 0.8307
Epoch 98/100
550/550 [==============================] - 0s 889us/step - loss: 0.3777 - accuracy: 0.8243 - val_loss: 0.3815 - val_accuracy: 0.8272
Epoch 99/100
550/550 [==============================] - 0s 905us/step - loss: 0.3770 - accuracy: 0.8250 - val_loss: 0.3757 - val_accuracy: 0.8288
Epoch 100/100
550/550 [==============================] - 1s 923us/step - loss: 0.3736 - accuracy: 0.8277 - val_loss: 0.3737 - val_accuracy: 0.8324
1717/1717 [==============================] - 2s 954us/step - loss: 0.3722 - accuracy: 0.8308
accuracy: 83.08%
```

Figure 4.28 - Model Execution Records

Visualization of the created Model shown in figure 4.29.

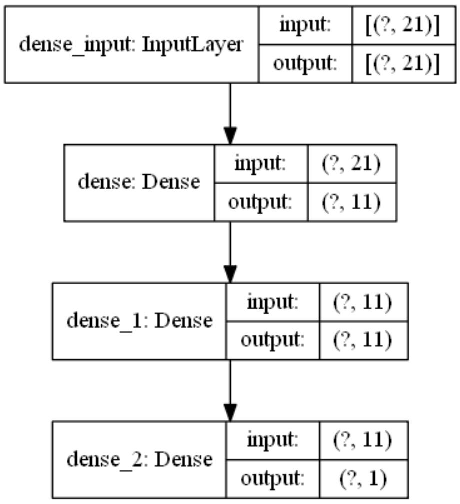
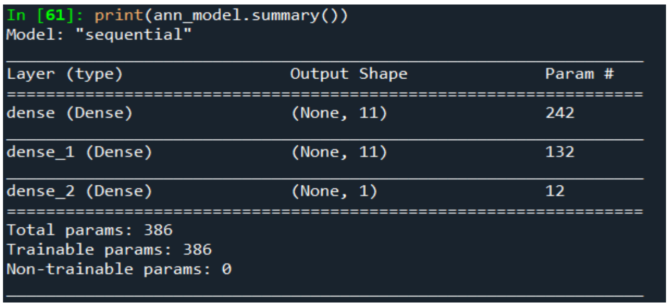


Figure 4.29 - ANN Model

We can see model summary of our model as shown in figure 4.30 . The figure shows layers and their order , output shape and number of parameters (weights) in each layers and total number of parameters (weight) in the model.

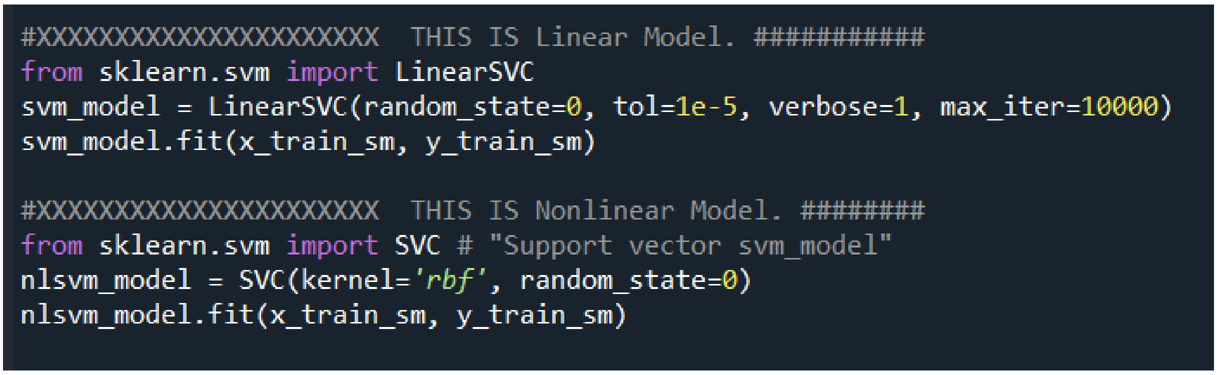


```
In [61]: print(ann_model.summary())
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 11)                242
_________________________________________________________________
dense_1 (Dense)              (None, 11)                132
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 12
=================================================================
Total params: 386
Trainable params: 386
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.30– ANN Model Summary

4.3.2 Application of Support Vector Machine

In this study we have applied linear and non-linear SVM. Linear and nonlinear model configuration shown in figure 4.31.



```
#XXXXXXXXXXXXXXXXXXXXXX  THIS IS Linear Model. ###########
from sklearn.svm import LinearSVC
svm_model = LinearSVC(random_state=0, tol=1e-5, verbose=1, max_iter=10000)
svm_model.fit(x_train_sm, y_train_sm)

#XXXXXXXXXXXXXXXXXXXXXX  THIS IS Nonlinear Model. ########
from sklearn.svm import SVC # "Support vector svm_model"
nlsvm_model = SVC(kernel='rbf', random_state=0)
nlsvm_model.fit(x_train_sm, y_train_sm)
```

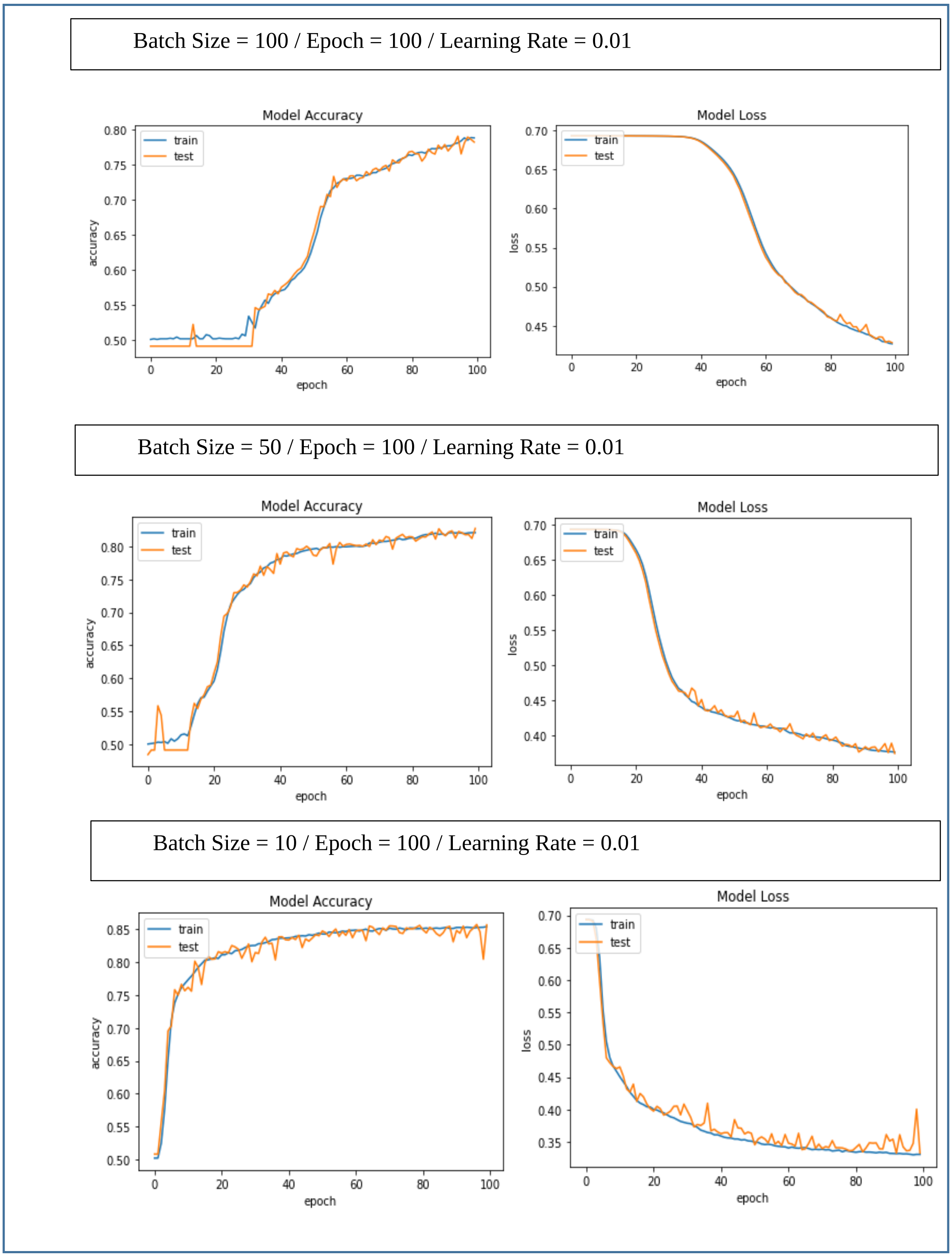
Figure 4.31- SVM Application

- LinearSVC – This is similar to SVC with parameter kernel=’linear’. But more flexible and scalable to large number of samples.
- Tol – This is tolerance for stopping criteria. Default is 1e-5
- Verbose - Enable verbose output.
- Max_iter - The maximum number of iterations to be run.
- Kernel – This is to specify the kernel type to be used in the algorithm. Default is Radial Basis Function (RBF)

4.4 Evaluation of ANN

Model Accuracy and Loss

We have validated our results by using different batch sizes and learning rates. Below figure 4.32 shown results of model execution for different batch sizes keeping learning rate at 0.01 (constant) and epoch = 100 (constant). Smaller batch generates high variance in the classification accuracy and loss. Smaller batch size provides slower learning process.



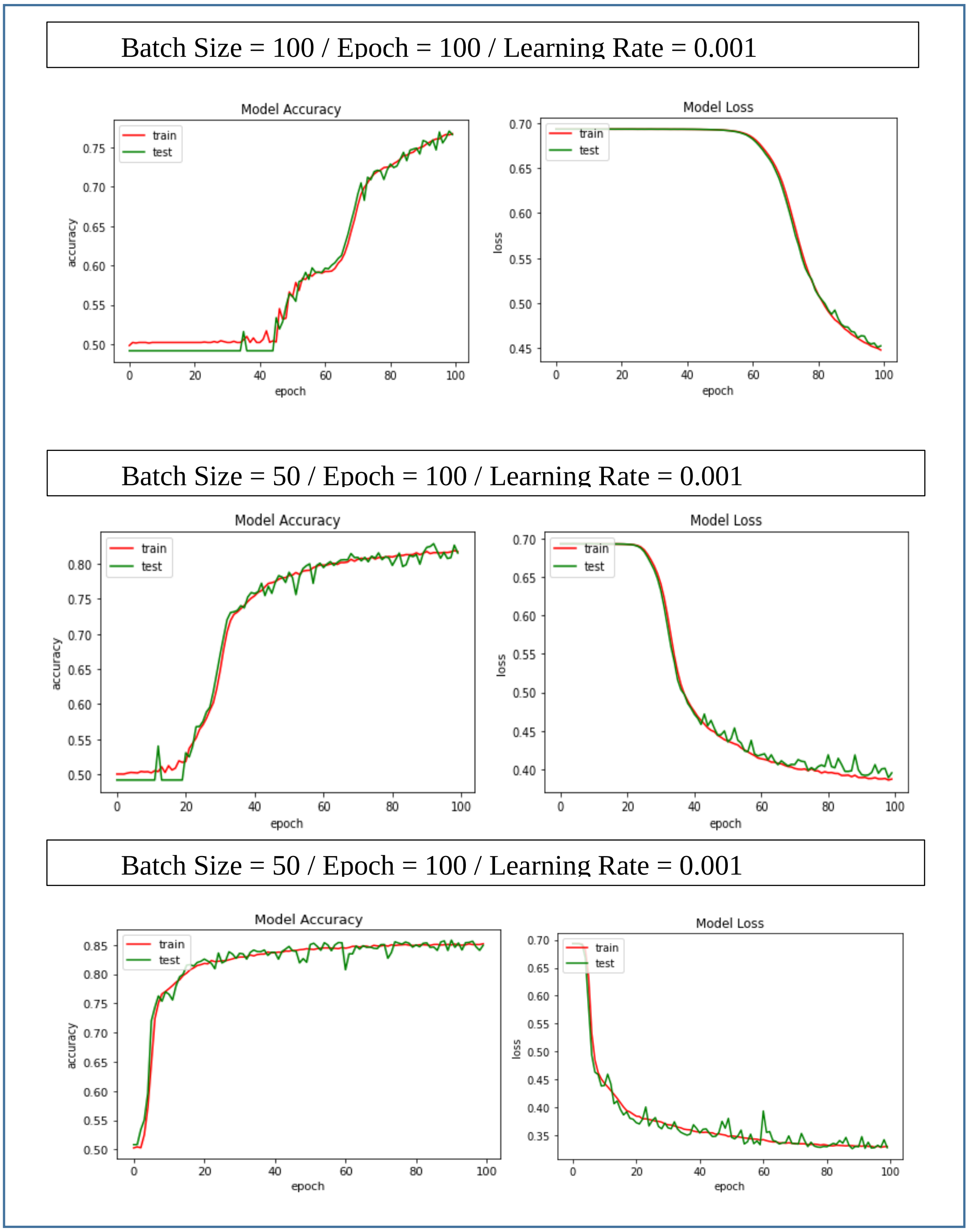
Batch Size = 100 / Epoch = 100 / Learning Rate = 0.01

Batch Size = 50 / Epoch = 100 / Learning Rate = 0.01

Batch Size = 10 / Epoch = 100 / Learning Rate = 0.01

Figure 4.32- Line plots of Accuracy and Loss with different batches – Large Learning Rate

Figure 4.33 shown results of model execution for different batch sizes keeping learning rate = 0.001 (constant) and epoch = 100 (constant). In here smaller learning rate used and accuracy is higher than using high learning rate. Smaller batch generates high variance in the classification accuracy and loss.



Batch Size = 100 / Epoch = 100 / Learning Rate = 0.001

Batch Size = 50 / Epoch = 100 / Learning Rate = 0.001

Batch Size = 50 / Epoch = 100 / Learning Rate = 0.001

Figure 4.33 - Line plots of Accuracy and Loss with different batches – Small Learning Rate

Model Accuracy, Model Loss, Model MAE (Mean Absolute Error) – High Learning Rate (0.01)

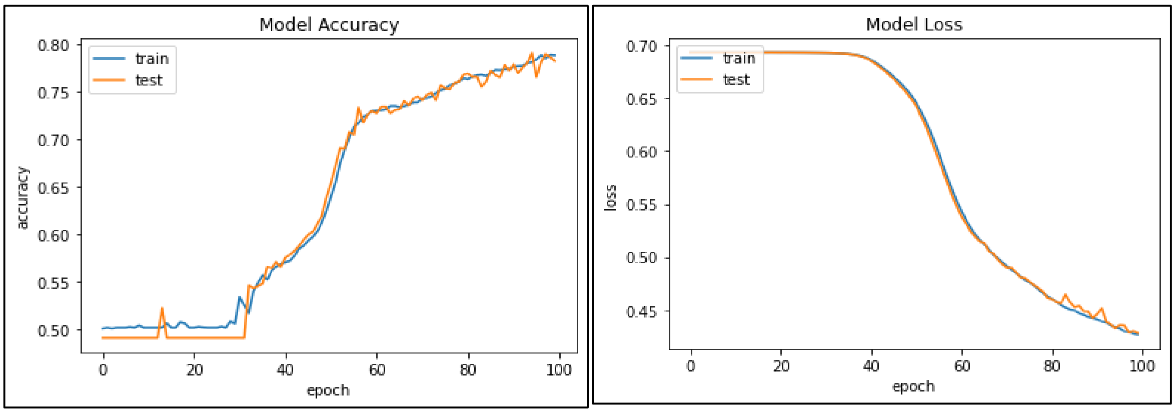


Figure 4.34 - Plot of Model Accuracy on Train and Validation Datasets

As shown in Figure 4.34 we can see trend for accuracy on training and test data set still rising and we can stop train the model. Furthermore, we can see trend for loss on training and test data set has comparable performance. If these parallel plots start moving further away from each other, it might be a sign to stop training.

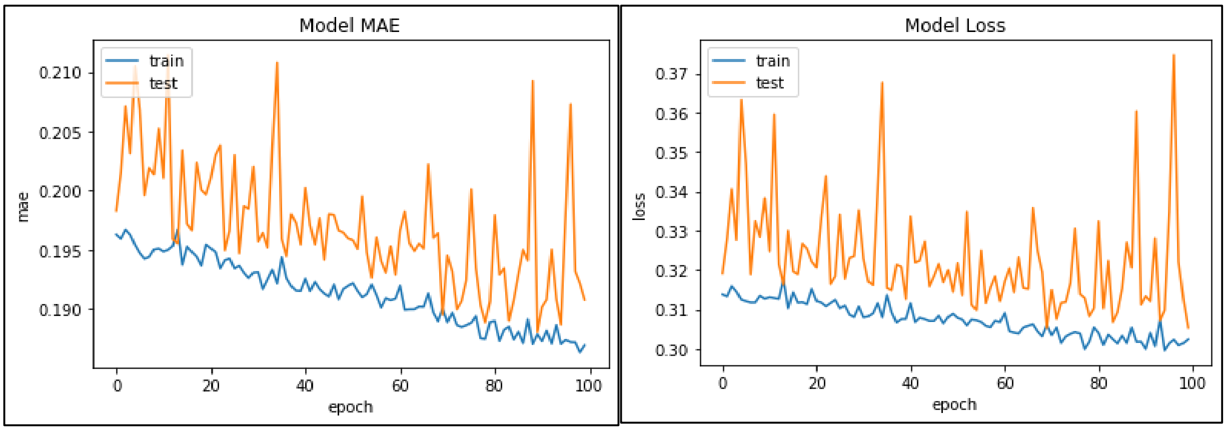


Figure 4.35 - Plot of Model MAE and Loss on Train and Validation Datasets

As shown in Figure 4.35 we can see trend for Mean Absolute error on training and test data set moving away from each other, still rising and we can train the model little further. We can see trend for loss on training and test data set has comparable performance. If these parallel plots start moving further away from each other, it might be a sign to stop training.

Confusion Matrix and Classification Report of Validation (Test) and Training Data – High Learning Rate (0.01)

The validation results we got for the predictions in validation and training data set by keeping parameters as epoch = 100, batch size = 100 and learning rate = 0.01 discussed below. Figure 4.36 shows confusion matrix. We can see false negative predictions are 1770 and false positive predictions are 1071 on testing data set. There are 6937 false negative predictions and 4096 false positive predictions in the training data set

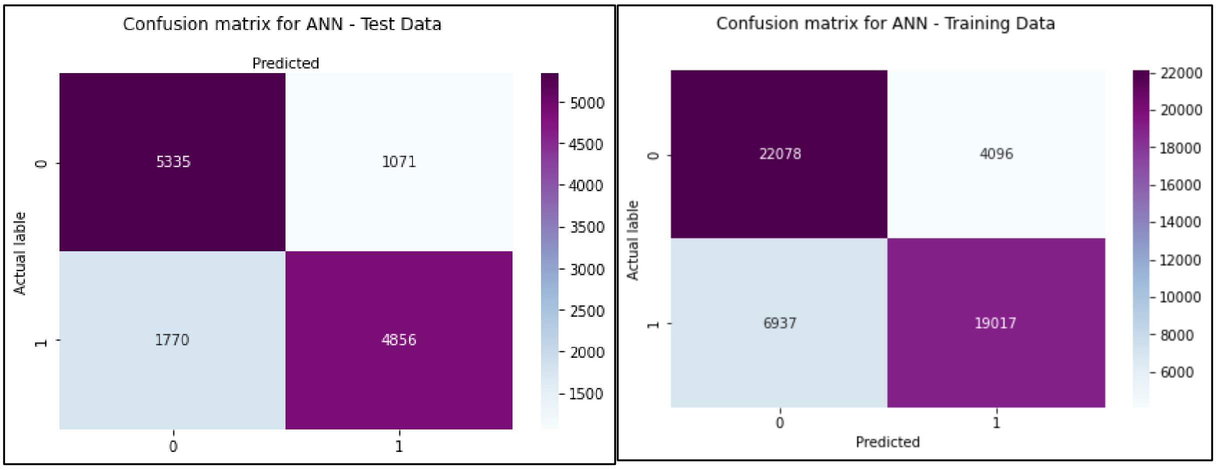
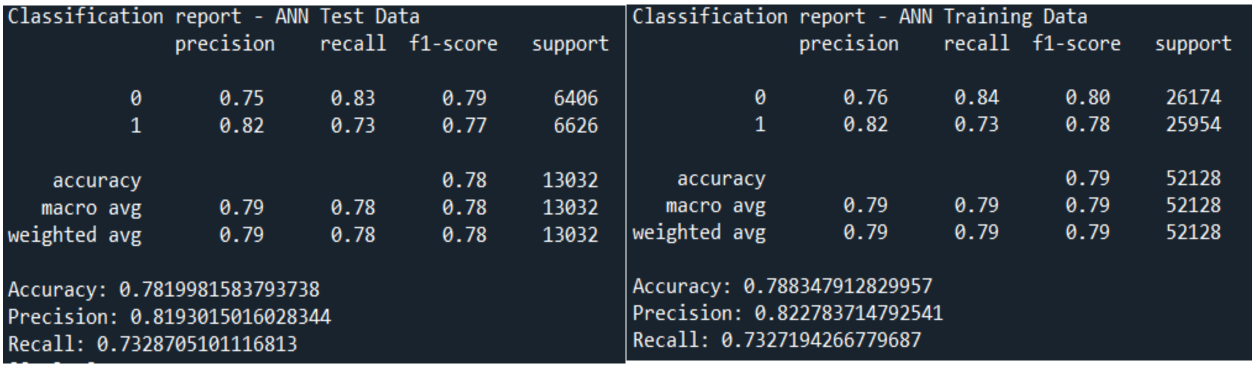


Figure 4.36 - Confusion Matrix on Validation and Training Data Set – ANN



```
Classification report - ANN Test Data
              precision    recall  f1-score   support

           0       0.75      0.83      0.79      6406
           1       0.82      0.73      0.77      6626

    accuracy                           0.78     13032
   macro avg       0.79      0.78      0.78     13032
weighted avg       0.79      0.78      0.78     13032

Accuracy: 0.7819981583793738
Precision: 0.8193015016028344
Recall: 0.7328705101116813
```

```
Classification report - ANN Training Data
              precision    recall  f1-score   support

           0       0.76      0.84      0.80     26174
           1       0.82      0.73      0.78     25954

    accuracy                           0.79     52128
   macro avg       0.79      0.79      0.79     52128
weighted avg       0.79      0.79      0.79     52128

Accuracy: 0.788347912829957
Precision: 0.822783714792541
Recall: 0.7327194266779687
```

Figure 4.37- Classification Report on Validation and Training Data Set – ANN

Figure 4.37 shows classification report of predictions in validation and testing data set. We can see Accuracy is 0.78, Precision is 0.81 and Recall is 0.73 in validation data set. In training data set Accuracy is 0.79, Precision is 0.82 and Recall is 0.73.

In our model validation accuracy and training accuracy are almost same.

Model Accuracy, Model Loss, Model MAE (Mean Absolute Error) – Low Learning Rate (0.001)

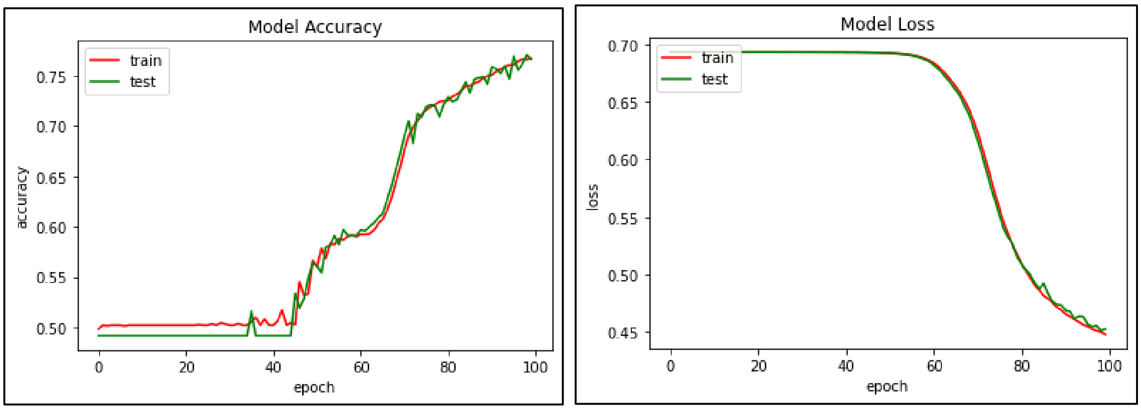


Figure 4.38 - Plot of Model Accuracy on Train and Validation Datasets

As shown in Figure 4.38 we can see trend for accuracy on training and test data set still rising and we can stop train the model. Furthermore, we can see trend for loss on training and test data set has comparable performance. If these parallel plots start moving further away from each other, it might be a sign to stop training.

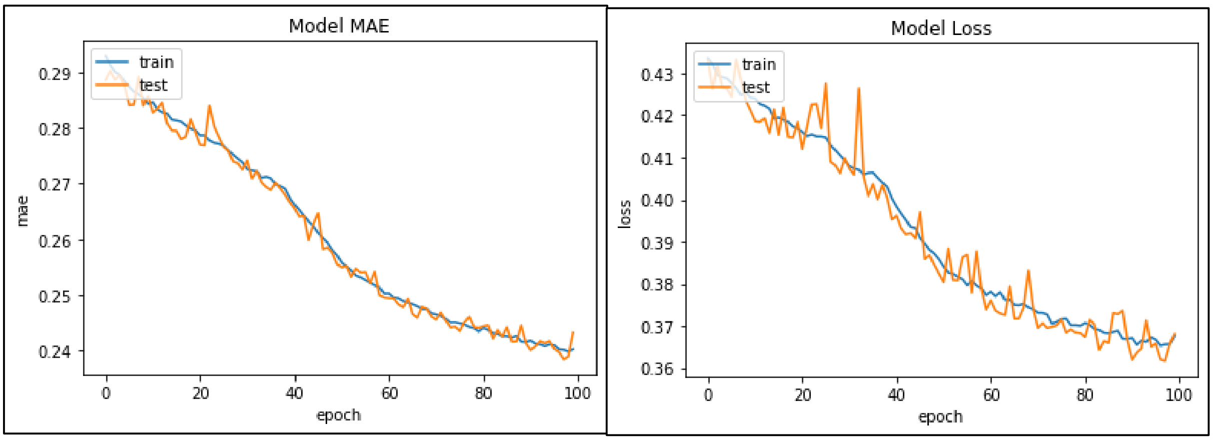


Figure 4.39 - Plot of Model MAE and Loss on Train and Validation Datasets

As shown in Figure 4.39 we can see trend for MAE on training and test data set still getting decreases and we can train the model little further. We can see trend for loss on training and test data set has comparable performance. If these parallel plots start moving further away from each other, it might be a sign to stop training.

Confusion Matrix and Classification Report of Validation (Test) and Training Data – High Learning Rate (0.001)

The validation results we got for the predictions in validation and training data set by keeping parameters as epoch = 100, batch size = 100 and learning rate = 0.001 discussed below. Figure 4.40 shows confusion matrix. We can see false negative predictions are 1770 and false positive predictions are 1400 on testing data set. There are 6844 false negative predictions and 5771 false positive predictions in training data set.

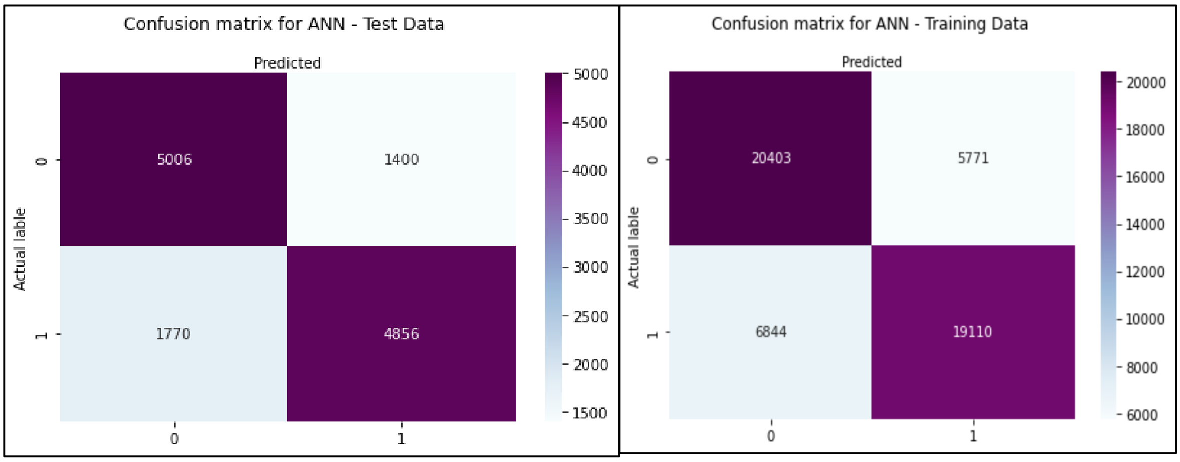
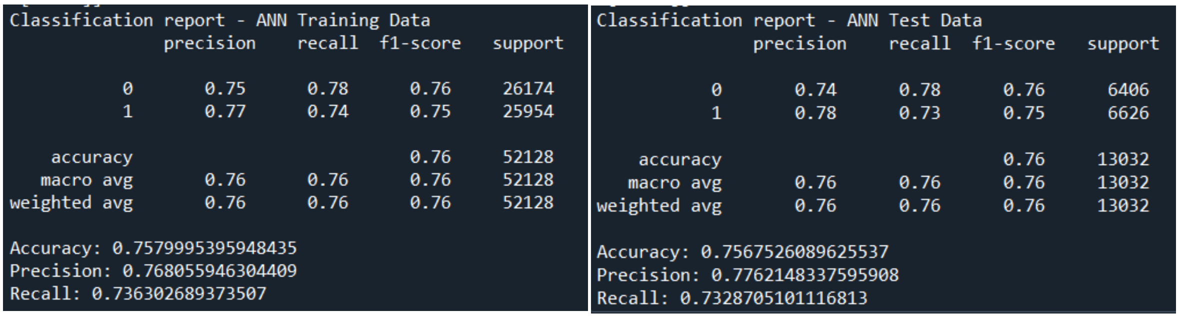


Figure 4.40 - Confusion Matrix on Validation and Training Data Set – ANN



```
Classification report - ANN Training Data
              precision    recall  f1-score   support

           0       0.75      0.78      0.76     26174
           1       0.77      0.74      0.75     25954

    accuracy                           0.76     52128
   macro avg       0.76      0.76      0.76     52128
weighted avg       0.76      0.76      0.76     52128

Accuracy: 0.7579995395948435
Precision: 0.768055946304409
Recall: 0.736302689373507
```

```
Classification report - ANN Test Data
              precision    recall  f1-score   support

           0       0.74      0.78      0.76      6406
           1       0.78      0.73      0.75      6626

    accuracy                           0.76     13032
   macro avg       0.76      0.76      0.76     13032
weighted avg       0.76      0.76      0.76     13032

Accuracy: 0.7567526089625537
Precision: 0.7762148337595908
Recall: 0.7328705101116813
```

Figure 4.41 - Classification Report on Validation and Training Data Set – ANN

Figure 4.41 shows classification report of predictions in validation and testing data set. We can see Accuracy is 0.76, Precision is 0.76 and Recall is 0.74 in validation data set. In training data set Accuracy is 0.76, Precision is 0.77 and Recall is 0.73. In our model validation accuracy and training accuracy are almost same.

Area Under the Curve (AUC) shown in figure 4.42 for validation and training data set under different learning rates.

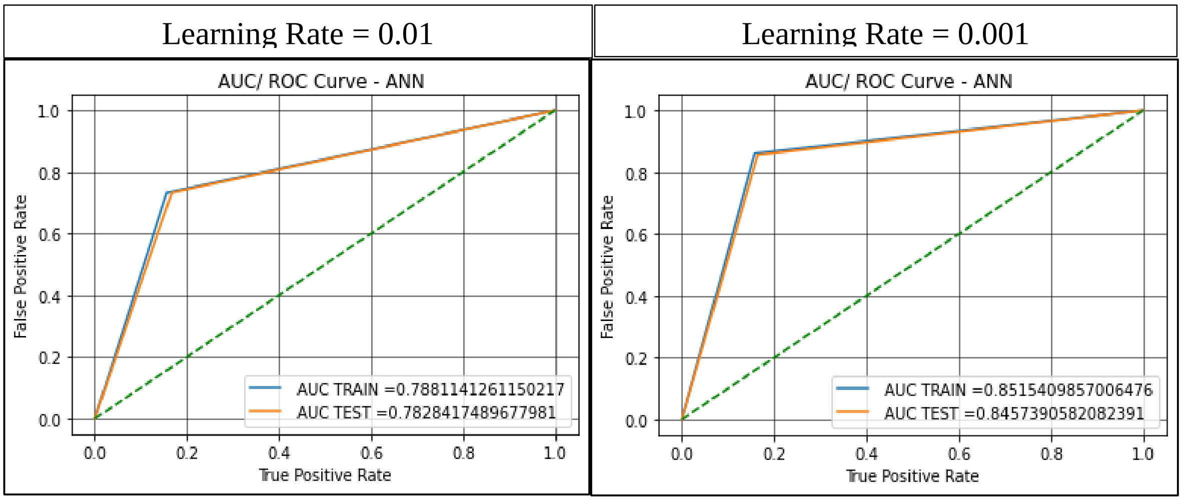


Figure 4.42 - ROC Curve of ANN

4.5 Evaluation of SVM

We have applied linear and Nonlinear models for our data set. Figure 4.43 Confusion Matrix shows validation results for linear SVM on testing and training data set. We can see false negative predictions are 755 and false positive predictions are 3086 in testing data set. In training data set there are 2979 false negative predictions and 11805 false positive predictions.

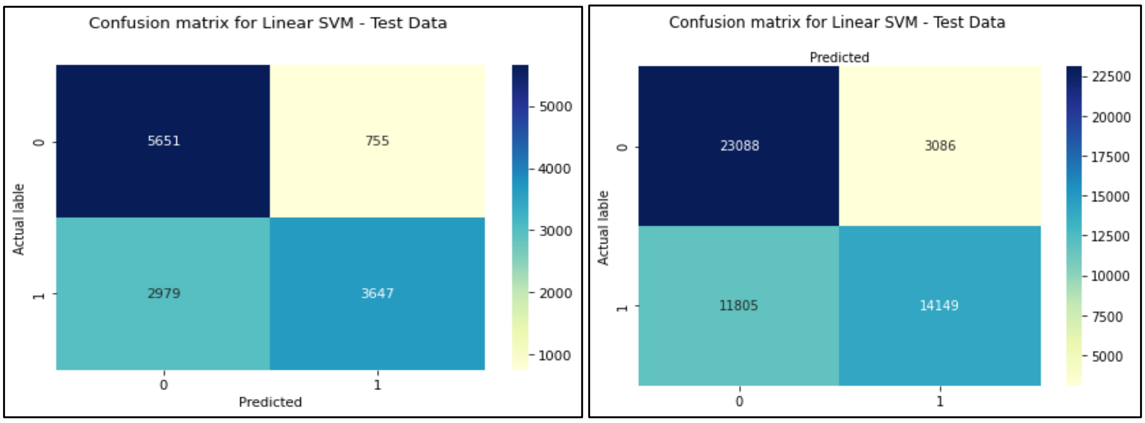
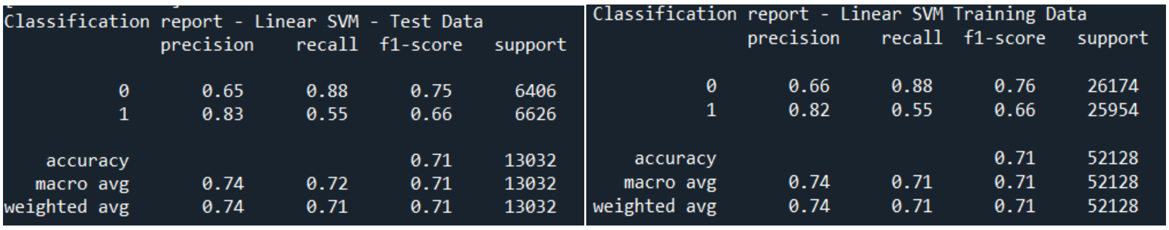


Figure 4.43 - Confusion Matrix on Validation and Training Data Set – Linear SVM



```
Classification report - Linear SVM - Test Data
              precision    recall  f1-score   support

           0       0.65      0.88      0.75      6406
           1       0.83      0.55      0.66      6626

    accuracy                           0.71     13032
   macro avg       0.74      0.72      0.71     13032
weighted avg       0.74      0.71      0.71     13032
```

```
Classification report - Linear SVM Training Data
              precision    recall  f1-score   support

           0       0.66      0.88      0.76     26174
           1       0.82      0.55      0.66     25954

    accuracy                           0.71     52128
   macro avg       0.74      0.71      0.71     52128
weighted avg       0.74      0.71      0.71     52128
```

Figure 4.44 - Classification Report on Validation and Training Data Set – Linear SVM

Figure 4.44 shows classification report of predictions in training data set. We can see Accuracy is 0.71, Precision is 0.83 and Recall is 0.55 in validation data set. In training data set Accuracy is 0.71, Precision is 0.82 and Recall is 0.55 in validation data set.

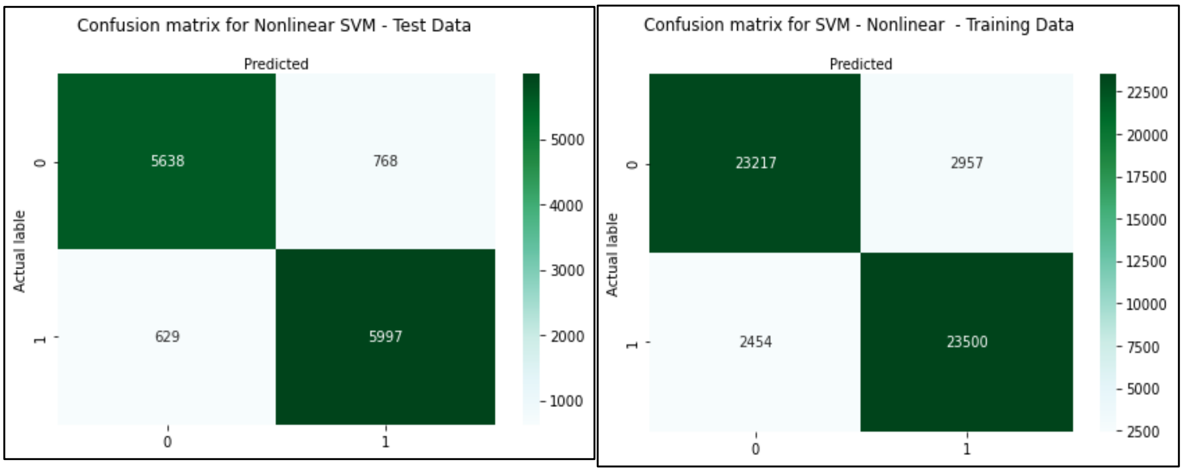
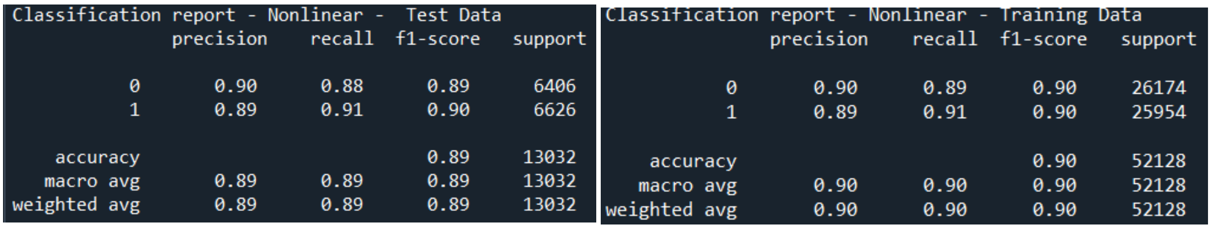


Figure 4.45 - Confusion Matrix on Validation and Training Data Set – Nonlinear SVM

Figure 4.45 Shown Confusion Matrixes for Nonlinear SVM on testing and training data set. There are 629 false negative predictions and 768 false positive predictions in testing data set. Furthermore, we can see there are 2454 false negative predictions and 23500 false positive predictions in in training data set.



```
Classification report - Nonlinear -  Test Data
              precision    recall  f1-score   support

           0       0.90      0.88      0.89      6406
           1       0.89      0.91      0.90      6626

    accuracy                           0.89     13032
   macro avg       0.89      0.89      0.89     13032
weighted avg       0.89      0.89      0.89     13032
```

```
Classification report - Nonlinear - Training Data
              precision    recall  f1-score   support

           0       0.90      0.89      0.90     26174
           1       0.89      0.91      0.90     25954

    accuracy                           0.90     52128
   macro avg       0.90      0.90      0.90     52128
weighted avg       0.90      0.90      0.90     52128
```

Figure 4.46 - Classification Report on Validation and Training Data Set – Nonlinear SVM

Figure 4.46 shows classification report of predictions in training and testing data set of Nonlinear SVM. We can see Accuracy is 0.88, Precision is 0.88 and Recall is 0.90 in validation data set. In training data set Accuracy is 0.89, Precision is 0.88 and Recall is 0.89. In our model validation accuracy and training accuracy are almost same.

Area Under the Curve (AUC) shown in figure 4.47 for validation and training data set for Nonlinear SVM.

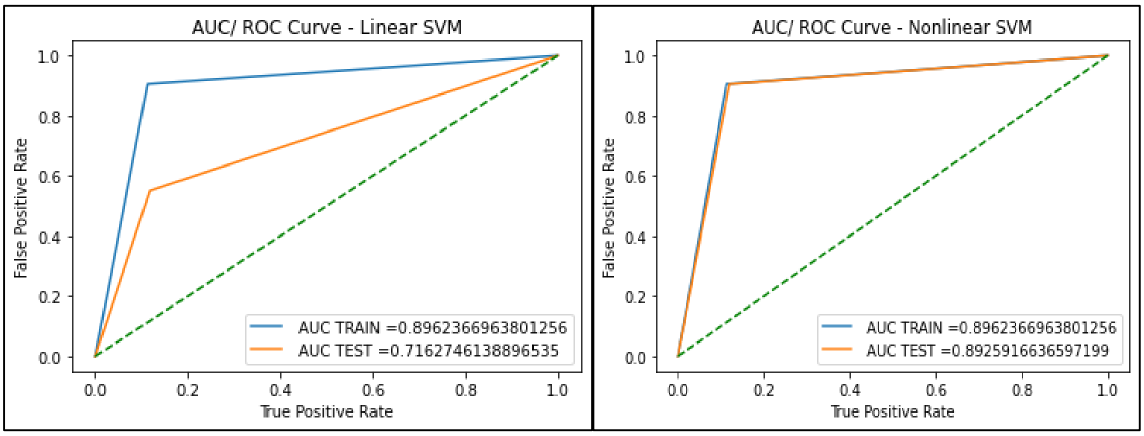


Figure 4.47 - ROC Curve on Linear and Nonlinear SVM

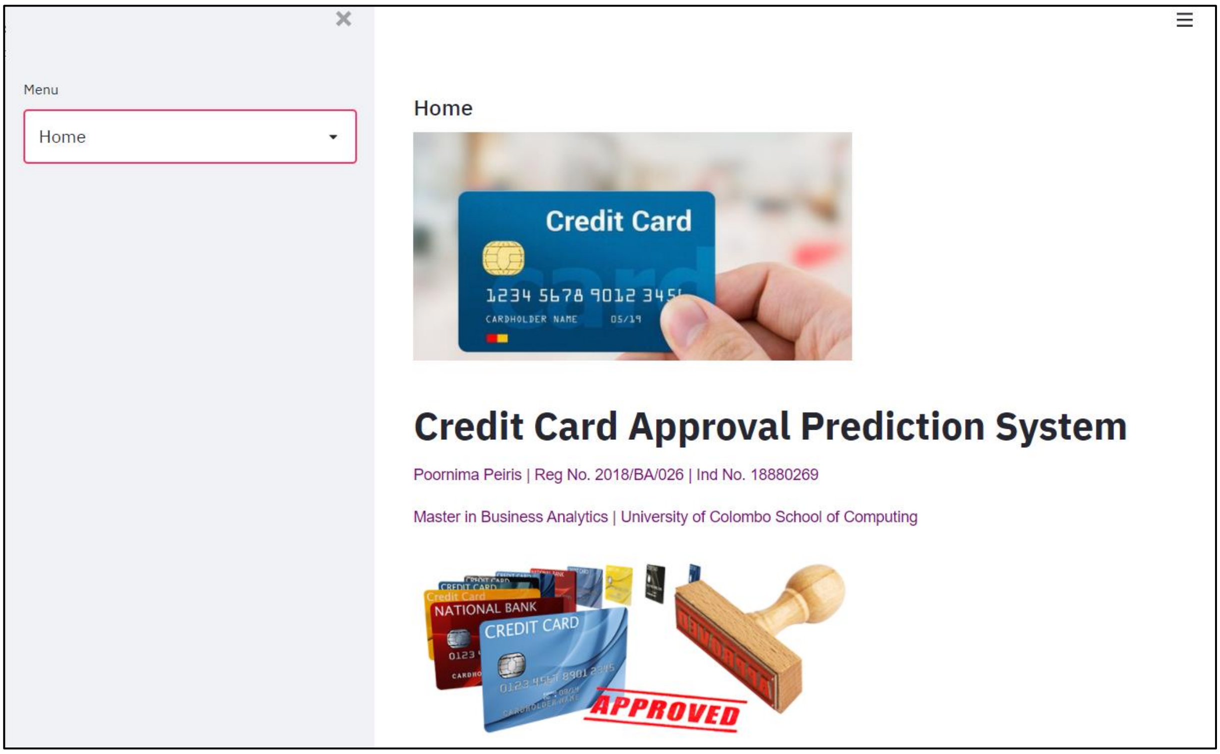
Table 4 5 - Summary of Model Accuracy

	ANN Learning Rate 0.01	**ANN Learning Rate 0.001**	**Linear SVM**	**Nonlinear SVM**
Accuracy	0.78	0.76	0.71	0.88
Precision	0.81	0.76	0.83	0.88
Recall	0.73	0.74	0.55	0.90
AUC/ ROC	0.79	0.85	0.89	0.89

Table 4.5 shows summary of model performance measurements under different evaluation criteria.

4.6 Deployment

After going through the model validation process, it was highlighted that Nonlinear SVM performs better than others. Therefore, we have decided to deploy the prediction model by using Nonlinear SVM. Firstly, we saved the nonlinear classification model to a pickle file. Here, we are saving our training model and will be using this for deploying the model. Use ‘streamlit’ for model deployment. Streamlit is an open-source python library which is easy to use and we can create beautiful web apps. There are two options. To predict a single customer entry and to predict a bulk set of data. Python script with streamlit created to implement the application. Samples of screen images of the application discuss in below. Main window of the app shown in figure 4.48.

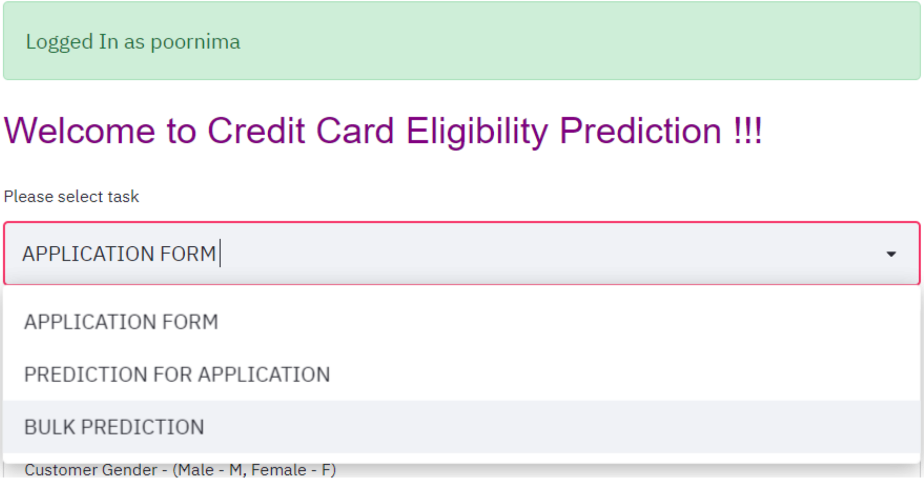


Credit Card Approval Prediction System

Poornima Peiris | Reg No. 2018/BA/026 | Ind No. 18880269

Master in Business Analytics | University of Colombo School of Computing

Figure 4.48 - Application Main Page



Logged In as poornima

Welcome to Credit Card Eligibility Prediction !!!

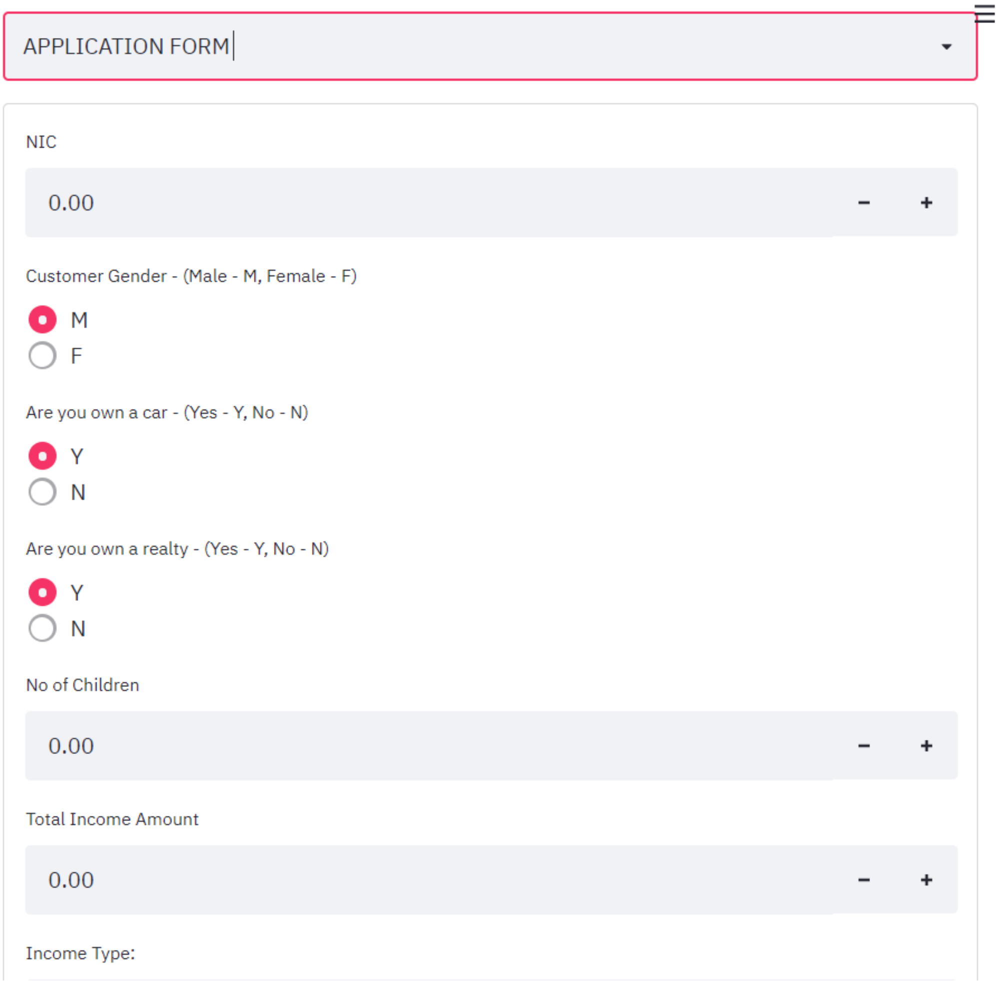
Please select task

APPLICATION FORM

APPLICATION FORM
PREDICTION FOR APPLICATION
BULK PREDICTION
Customer Gender - (Male - M, Female - F)

Figure 4.49 - Application Selection Menu

After logging into the system there are three options to select. First enter individual customer level data by using the ‘application form’ tab. Next is to get the prediction by using the ‘prediction for application’ tab as shown in figure 4.49. Applicant data can be entering into the provided form as shown in figure 4.50



APPLICATION FORM

NIC

0.00

Customer Gender - (Male - M, Female - F)

● M
○ F

Are you own a car - (Yes - Y, No - N)

● Y
○ N

Are you own a realty - (Yes - Y, No - N)

● Y
○ N

No of Children

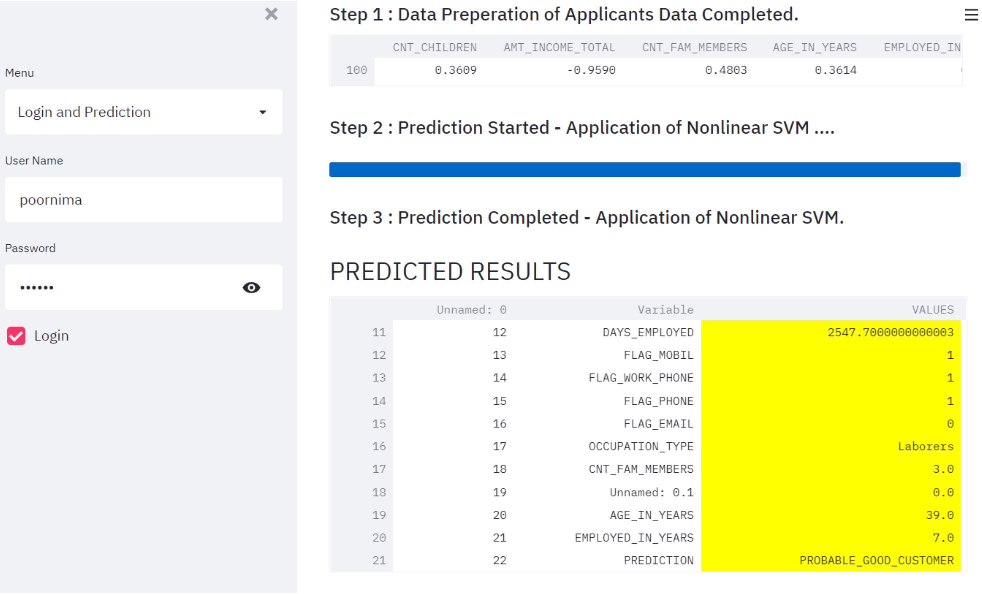
0.00

Total Income Amount

0.00

Income Type:

Figure 4.50 - Sample Application form



User Name: poornima

Password: ••••••

☑ Login

Step 1 : Data Preperation of Applicants Data Completed.

	CNT_CHILDREN	AMT_INCOME_TOTAL	CNT_FAM_MEMBERS	AGE_IN_YEARS	EMPLOYED_IN
100	0.3609	-0.9590	0.4803	0.3614	

Step 2 : Prediction Started - Application of Nonlinear SVM ....

Step 3 : Prediction Completed - Application of Nonlinear SVM.

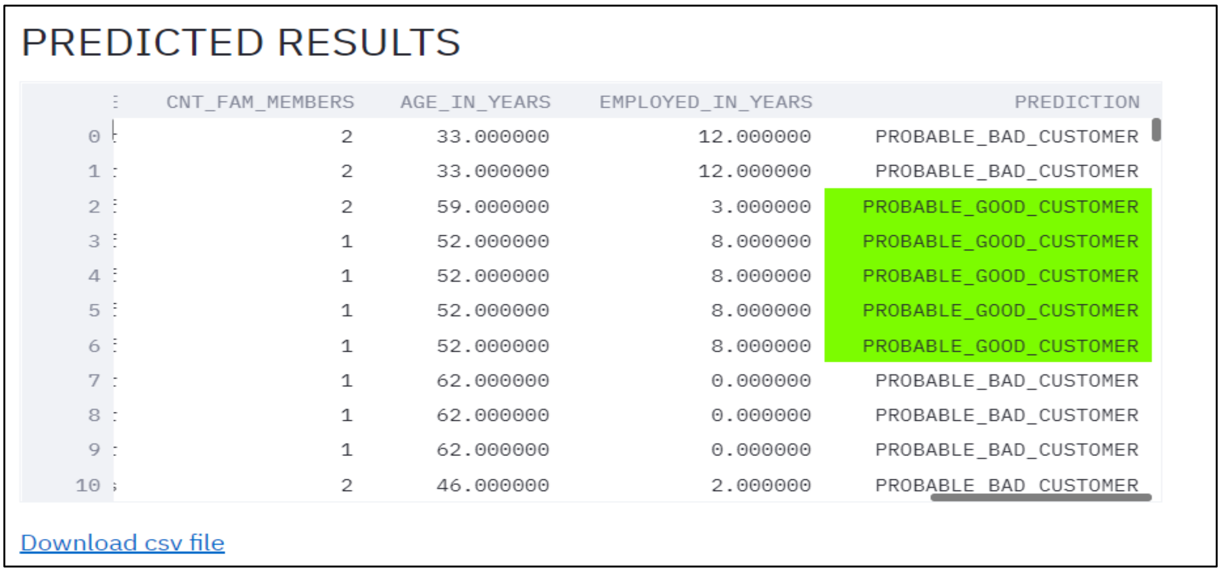
PREDICTED RESULTS

	Unnamed: 0	Variable	VALUES
11	12	DAYS_EMPLOYED	2547.7000000000003
12	13	FLAG_MOBIL	1
13	14	FLAG_WORK_PHONE	1
14	15	FLAG_PHONE	1
15	16	FLAG_EMAIL	0
16	17	OCCUPATION_TYPE	Laborers
17	18	CNT_FAM_MEMBERS	3.0
18	19	Unnamed: 0.1	0.0
19	20	AGE_IN_YEARS	39.0
20	21	EMPLOYED_IN_YEARS	7.0
21	22	PREDICTION	PROBABLE_GOOD_CUSTOMER

Figure 4.51 - Sample Application Form Prediction Result

Predicted result for Non-linear SVM application shown in figure 4.51

Furthermore, the set of application data can be predicted as bulk by using 3rd option in the menu. Sample of bulk prediction results is shown in figure 4.52.



PREDICTED RESULTS

	CNT_FAM_MEMBERS	AGE_IN_YEARS	EMPLOYED_IN_YEARS	PREDICTION
0	2	33.000000	12.000000	PROBABLE_BAD_CUSTOMER
1	2	33.000000	12.000000	PROBABLE_BAD_CUSTOMER
2	2	59.000000	3.000000	PROBABLE_GOOD_CUSTOMER
3	1	52.000000	8.000000	PROBABLE_GOOD_CUSTOMER
4	1	52.000000	8.000000	PROBABLE_GOOD_CUSTOMER
5	1	52.000000	8.000000	PROBABLE_GOOD_CUSTOMER
6	1	52.000000	8.000000	PROBABLE_GOOD_CUSTOMER
7	1	62.000000	0.000000	PROBABLE_BAD_CUSTOMER
8	1	62.000000	0.000000	PROBABLE_BAD_CUSTOMER
9	1	62.000000	0.000000	PROBABLE_BAD_CUSTOMER
10	2	46.000000	2.000000	PROBABLE_BAD_CUSTOMER

Download csv file

Figure 4.52 - Sample Predicted Results

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

We have obtained the publically available data set and explanatory analysis was carried out to understand the data set. Then conducted several activities related to data preparations such as data preprocessing, feature selections and feature scaling. To achieve a desired outcome, it is very important to carry out these activities accurately. We have divided the data set into two parts as a training and test data set and the intended purpose is to validate the accuracy of the model. Artificial Neural Network, Linear SVM and Nonlinear SVM three predictive models were implemented. Performance measures were tested by using Accuracy, Precision, Recall, AUC on each classifier. In ANN Mean Absolute Error (MAE) is tested.

ANN model performances tested using low and high learning rates. Accuracy is 0.78, Precision is 0.81, Recall is 0.73 and AUC is 0.79 with a higher learning rate of 0.01. Accuracy is 0.76, Precision is 0.76, Recall is 0.74 and AUC is 0.85 with a lower learning rate of 0.001. Precision and recall values are high in higher learning rate. Smaller batch size provides a slower learning process. However, a small learning rate gave better AUC at 0.85 for ANN compared to high learning rate.

In the linear SVM model, Accuracy is 0.71, Precision is 0.83, Recall is 0.55 and AUC is 0.89. Here we can see the recall is low compared to ANN. However, in nonlinear SVM model Accuracy is 0.88, Precision is 0.88, Recall is 0.90 and AUC is 0.89. Accuracy, Precision and Recall values are higher in Nonlinear SVM than ANN and Linear SVM. Recall rate is 0.90 means the model predicts positive class 90% correctly.

We have evaluated three classifiers and observed that Nonlinear SVM is performed better than ANN and linear SVM. We have achieved a high accuracy level and by considering that usage of this model in the real world can be applicable. This data set included main demographic data relevant to Sri Lankan context. Hence, application of local context can be considered.

5.2 Future Work

We realized that customer behavior might be different country to country and application of several real banking datasets can be considered for further studies. To consider default customers not only their demographic and socio-cultural data but also other existing credit facilities information such as other loans can be taken as features to get more accurate results. This data set was generated before the pandemic situation. During the current pandemic situation of COVID-19 there is an increase of defaulters and their paying behaviors are different than before the pandemic. Economic conditions have changed due to the pandemic. Application of data sets including COVID-19 impact under new normal to be an area of concern for researchers. Furthermore, the data set is a highly imbalanced data set and we have applied SMORTE for balancing. Whether there is a relationship between Nonlinearity in highly imbalanced class problems with SMORTE application is another area of concern for researchers.

APPENDICES

A. Tools and Technology

We have implement this project by using Microsoft Power BI for Data Visualization and Python with Spider (as Scientific Python Development Environment) for data programing. There were several python libraries being used for analyzing the data, model building and model validation. They are Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn, TensorFlow, imblearn, keras, joblib and streamlit.

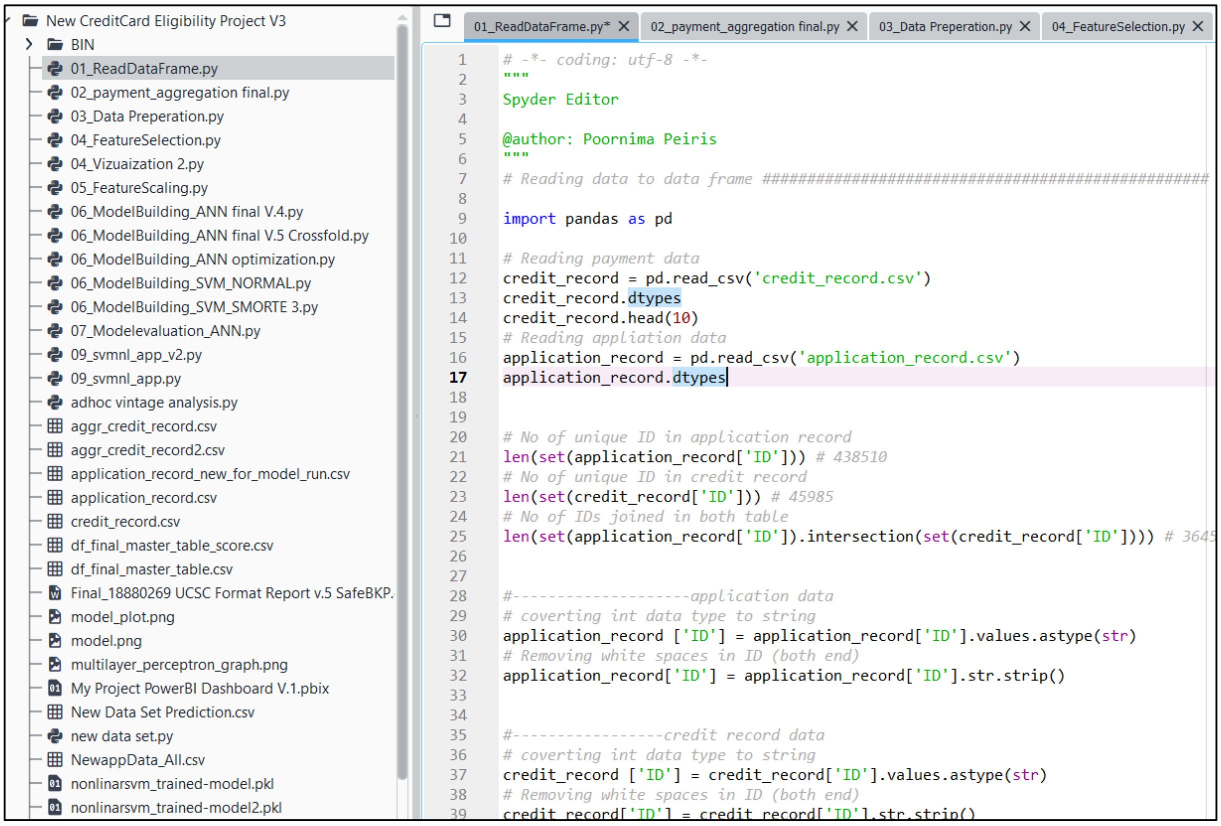
Table – Usage of libraries

Pandas	Mainly used for data analysis, import data from csv files, data manipulation such as selecting, merging, reshaping and for data cleaning process.
NumPy	Used for working with n-dimensional arrays and linear algebra
Matplotlib	Used for creating static and interactive visualization in data
Seaborn	Used for creating advanced visualization in data
Scikit-Learn	Used for implementing machine learning algorithms, divide data set as training and test, data encoding and model validation such as confusion matrix. • category_encoders • StandardScaler • model_selection import train_test_split • import metrics • metrics import classification_report, confusion_matrix • LinearSVC
TensorFlow	Used for implementing machine learning algorithm
Imblearn	Used for generating synthetic minority over-sampling technique (SMORTE) • imblearn.over_sampling import SMOTE
Keras	Used for implementing ANN machine learning algorithm • from keras.models import Sequential • from keras.layers import Dense
joblib	Running Python functions as pipeline jobs. Used to save and open the model.
Streamlit	Streamlit is an open-source python library which is easy to use and can create beautiful web apps. Used for model deployment

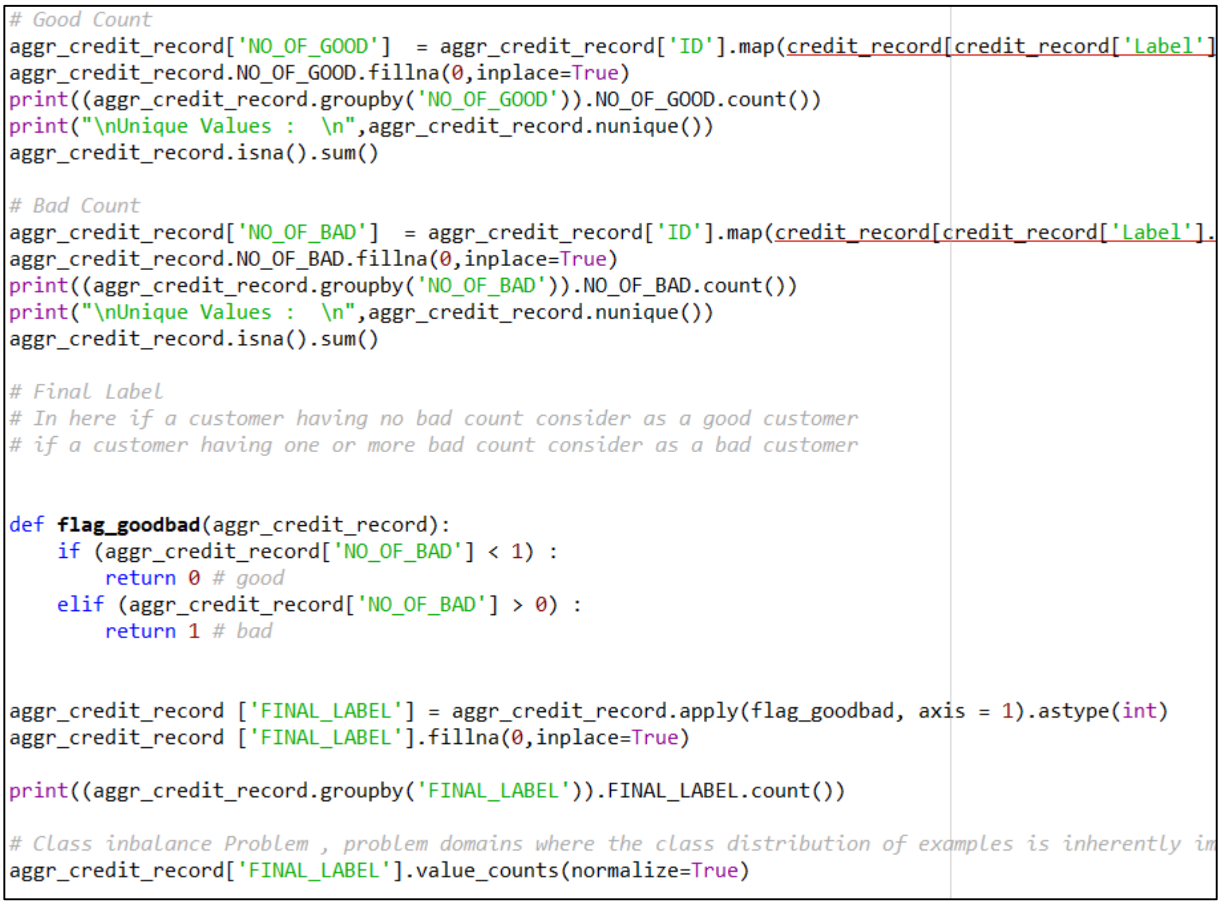
B. List of Categorical Features with Information Value

Feature Name	Value
FLAG_WORK_PHONE_1	0.007271
FLAG_EMAIL_1	0.006114
NAME_HOUSING_TYPE_House apartment	0.005997
NAME_FAMILY_STATUS_Married	0.004869
CODE_GENDER_0	0.004784
FLAG_PHONE_1	0.004142
FLAG_OWN_CAR_0	0.003852
NAME_EDUCATION_TYPE	0.003109
FLAG_OWN_REALTY_1	0.003042
NAME_INCOME_TYPE_Working	0.002278
FLAG_PHONE_0	0.002056
FLAG_OWN_REALTY_0	0.001848
FLAG_OWN_CAR_1	0.001744
OCCUPATION_TYPE_IT staff	0.001161
CODE_GENDER_1	0.001133
OCCUPATION_TYPE_Accountants	0.001018
OCCUPATION_TYPE_Other	0.000491
FLAG_EMAIL_0	0.000464
NAME_HOUSING_TYPE_With parents	0.000453
NAME_INCOME_TYPE_Commercial_associate	0.000408
OCCUPATION_TYPE_Cleaning staff	0.000376
OCCUPATION_TYPE_Medicine Staff	0.000354
OCCUPATION_TYPE_Cooking staff	0.000322
FLAG_WORK_PHONE_0	0.000294
OCCUPATION_TYPE_Private service staff	0.000216
NAME_INCOME_TYPE_State_servant	0.000182
NAME_HOUSING_TYPE_Rented apartment	0.000071
OCCUPATION_TYPE_Waiters/barmen staff	0.000045
OCCUPATION_TYPE_Sales Staff	0.000004
NAME_INCOME_TYPE_Pensioner	0.000000
NAME_INCOME_TYPE_Student	0.000000
NAME_FAMILY_STATUS_Civil marriage	0.000000
NAME_FAMILY_STATUS_Separated	0.000000
NAME_FAMILY_STATUS_Single not married	0.000000
NAME_FAMILY_STATUS_Widow	0.000000
NAME_HOUSING_TYPE_Co-op apartment apartment	0.000000
OCCUPATION_TYPE_Core staff	0.000000
OCCUPATION_TYPE_Drivers	0.000000
OCCUPATION_TYPE_HR staff	0.000000
OCCUPATION_TYPE_High skill tech staff	0.000000
OCCUPATION_TYPE_Laborers	0.000000
OCCUPATION_TYPE_Low-skill Laborers	0.000000
OCCUPATION_TYPE_Managers	0.000000
OCCUPATION_TYPE_Realty Agents	0.000000
OCCUPATION_TYPE_Secretaries	0.000000
OCCUPATION_TYPE_Security staff	0.000000
NAME_HOUSING_TYPE_Municipal apartment	0.000000
NAME_HOUSING_TYPE_Office apartment	0.000000

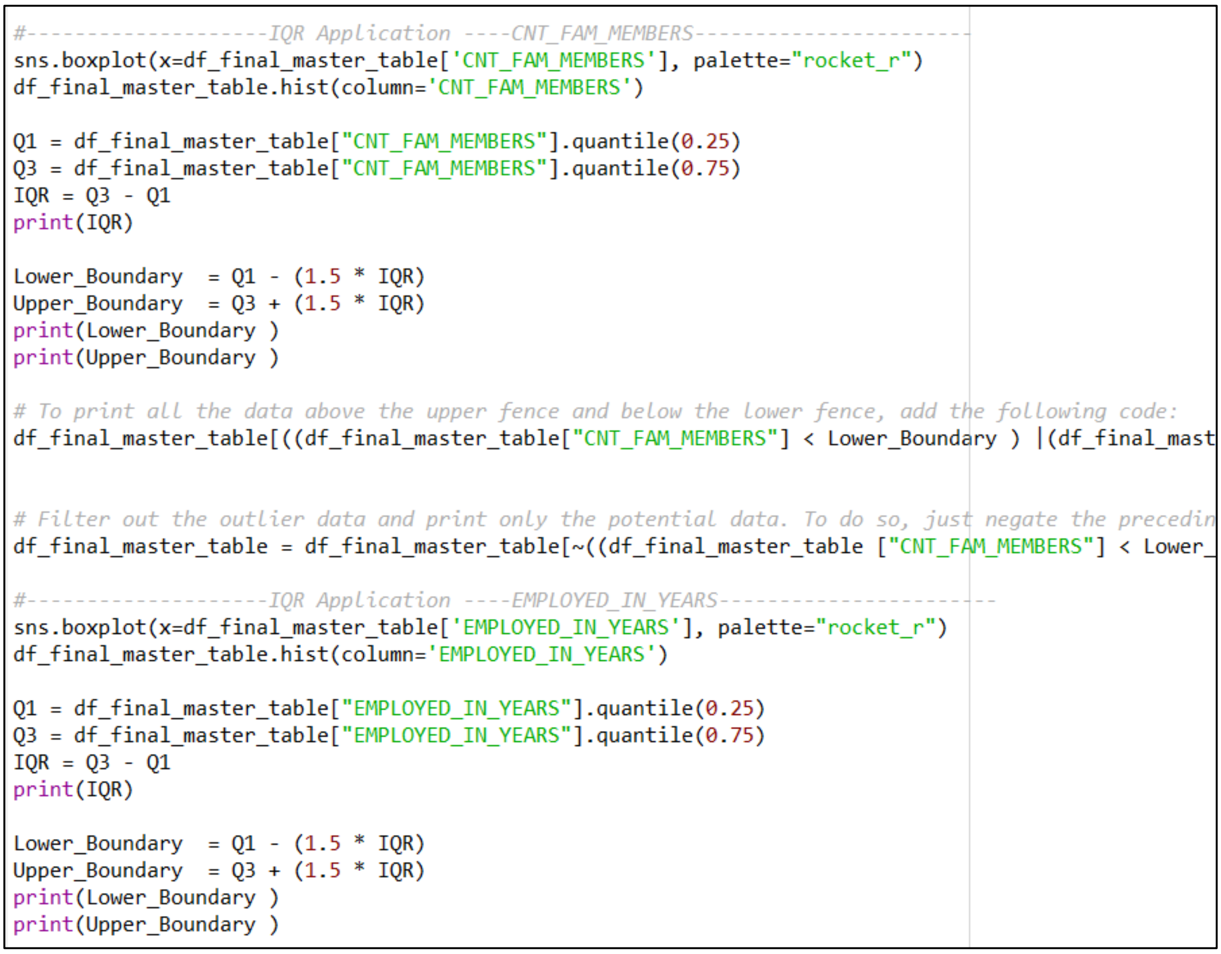
C. Code Samples


New CreditCard Eligibility Project V3
BIN
01_ReadDataFrame.py
02_payment_aggregation final.py
03_Data Preperation.py
04_FeatureSelection.py
04_Vizuaization 2.py
05_FeatureScaling.py
06_ModelBuilding_ANN final V.4.py
06_ModelBuilding_ANN final V.5 Crossfold.py
06_ModelBuilding_ANN optimization.py
06_ModelBuilding_SVM_NORMAL.py
06_ModelBuilding_SVM_SMORTE 3.py
07_Modelevaluation_ANN.py
09_svmnl_app_v2.py
09_svmnl_app.py
adhoc vintage analysis.py
aggr_credit_record.csv
aggr_credit_record2.csv
application_record_new_for_model_run.csv
application_record.csv
credit_record.csv
df_final_master_table_score.csv
df_final_master_table.csv
Final_18880269 UCSC Format Report v.5 SafeBKP.
model_plot.png
model.png
multilayer_perceptron_graph.png
My Project PowerBI Dashboard V.1.pbix
New Data Set Prediction.csv
new data set.py
NewappData_All.csv
nonlinarsvm_trained-model.pkl
nonlinarsvm_trained-model2.pkl
01_ReadDataFrame.py*
02_payment_aggregation final.py
03_Data Preperation.py
04_FeatureSelection.py

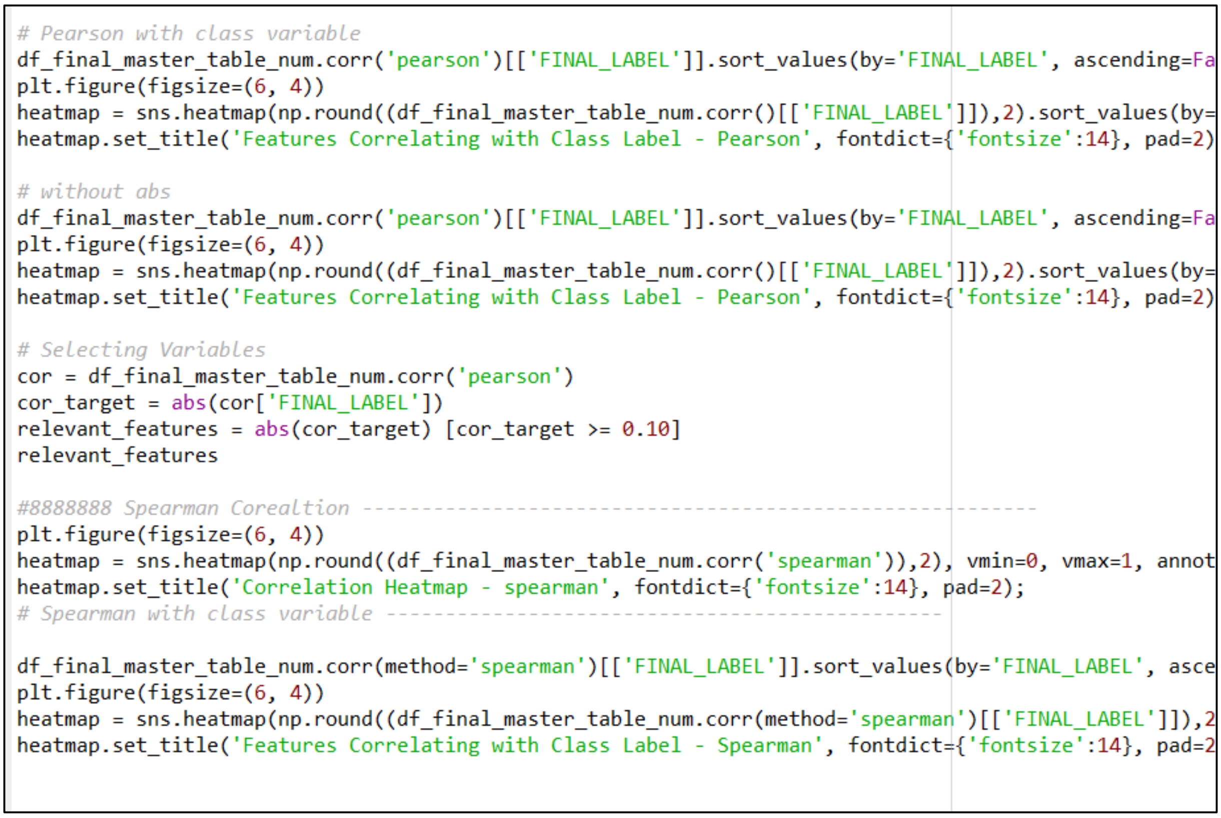
```
# -*- coding: utf-8 -*-
"""
Spyder Editor

@author: Poornima Peiris
"""
# Reading data to data frame ##################################################

import pandas as pd

# Reading payment data
credit_record = pd.read_csv('credit_record.csv')
credit_record.dtypes
credit_record.head(10)
# Reading appliation data
application_record = pd.read_csv('application_record.csv')
application_record.dtypes



# No of unique ID in application record
len(set(application_record['ID'])) # 438510
# No of unique ID in credit record
len(set(credit_record['ID'])) # 45985
# No of IDs joined in both table
len(set(application_record['ID']).intersection(set(credit_record['ID']))) # 3645


#-------------------application data
# coverting int data type to string
application_record ['ID'] = application_record['ID'].values.astype(str)
# Removing white spaces in ID (both end)
application_record['ID'] = application_record['ID'].str.strip()


#----------------credit record data
# coverting int data type to string
credit_record ['ID'] = credit_record['ID'].values.astype(str)
# Removing white spaces in ID (both end)
credit_record['ID'] = credit_record['ID'].str.strip()
```



```
# Good Count
aggr_credit_record['NO_OF_GOOD']  = aggr_credit_record['ID'].map(credit_record[credit_record['Label']
aggr_credit_record.NO_OF_GOOD.fillna(0,inplace=True)
print((aggr_credit_record.groupby('NO_OF_GOOD')).NO_OF_GOOD.count())
print("\nUnique Values :  \n",aggr_credit_record.nunique())
aggr_credit_record.isna().sum()

# Bad Count
aggr_credit_record['NO_OF_BAD']  = aggr_credit_record['ID'].map(credit_record[credit_record['Label'].
aggr_credit_record.NO_OF_BAD.fillna(0,inplace=True)
print((aggr_credit_record.groupby('NO_OF_BAD')).NO_OF_BAD.count())
print("\nUnique Values :  \n",aggr_credit_record.nunique())
aggr_credit_record.isna().sum()

# Final Label
# In here if a customer having no bad count consider as a good customer
# if a customer having one or more bad count consider as a bad customer


def flag_goodbad(aggr_credit_record):
    if (aggr_credit_record['NO_OF_BAD'] < 1) :
        return 0 # good
    elif (aggr_credit_record['NO_OF_BAD'] > 0) :
        return 1 # bad


aggr_credit_record ['FINAL_LABEL'] = aggr_credit_record.apply(flag_goodbad, axis = 1).astype(int)
aggr_credit_record ['FINAL_LABEL'].fillna(0,inplace=True)

print((aggr_credit_record.groupby('FINAL_LABEL')).FINAL_LABEL.count())

# Class inbalance Problem , problem domains where the class distribution of examples is inherently im
aggr_credit_record['FINAL_LABEL'].value_counts(normalize=True)
```

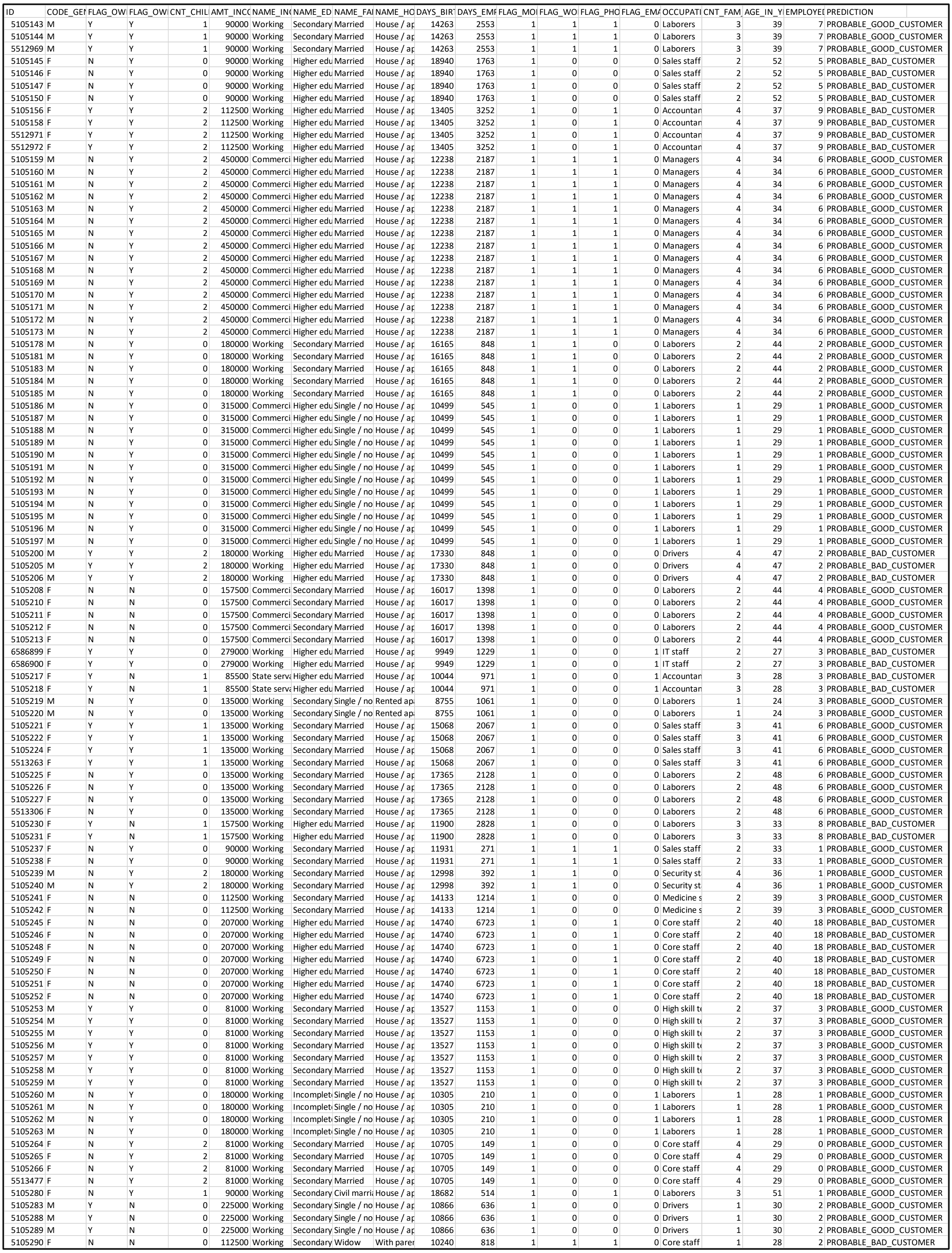


```
#--------------------IQR Application ----CNT_FAM_MEMBERS----------------------
sns.boxplot(x=df_final_master_table['CNT_FAM_MEMBERS'], palette="rocket_r")
df_final_master_table.hist(column='CNT_FAM_MEMBERS')

Q1 = df_final_master_table["CNT_FAM_MEMBERS"].quantile(0.25)
Q3 = df_final_master_table["CNT_FAM_MEMBERS"].quantile(0.75)
IQR = Q3 - Q1
print(IQR)

Lower_Boundary  = Q1 - (1.5 * IQR)
Upper_Boundary  = Q3 + (1.5 * IQR)
print(Lower_Boundary )
print(Upper_Boundary )

# To print all the data above the upper fence and below the lower fence, add the following code:
df_final_master_table[((df_final_master_table["CNT_FAM_MEMBERS"] < Lower_Boundary ) |(df_final_mast


# Filter out the outlier data and print only the potential data. To do so, just negate the precedin
df_final_master_table = df_final_master_table[~((df_final_master_table ["CNT_FAM_MEMBERS"] < Lower_

#--------------------IQR Application ----EMPLOYED_IN_YEARS----------------------
sns.boxplot(x=df_final_master_table['EMPLOYED_IN_YEARS'], palette="rocket_r")
df_final_master_table.hist(column='EMPLOYED_IN_YEARS')

Q1 = df_final_master_table["EMPLOYED_IN_YEARS"].quantile(0.25)
Q3 = df_final_master_table["EMPLOYED_IN_YEARS"].quantile(0.75)
IQR = Q3 - Q1
print(IQR)

Lower_Boundary  = Q1 - (1.5 * IQR)
Upper_Boundary  = Q3 + (1.5 * IQR)
print(Lower_Boundary )
print(Upper_Boundary )
```



```
# Pearson with class variable
df_final_master_table_num.corr('pearson')[['FINAL_LABEL']].sort_values(by='FINAL_LABEL', ascending=Fa
plt.figure(figsize=(6, 4))
heatmap = sns.heatmap(np.round((df_final_master_table_num.corr()[['FINAL_LABEL']]),2).sort_values(by=
heatmap.set_title('Features Correlating with Class Label - Pearson', fontdict={'fontsize':14}, pad=2)

# without abs
df_final_master_table_num.corr('pearson')[['FINAL_LABEL']].sort_values(by='FINAL_LABEL', ascending=Fa
plt.figure(figsize=(6, 4))
heatmap = sns.heatmap(np.round((df_final_master_table_num.corr()[['FINAL_LABEL']]),2).sort_values(by=
heatmap.set_title('Features Correlating with Class Label - Pearson', fontdict={'fontsize':14}, pad=2)

# Selecting Variables
cor = df_final_master_table_num.corr('pearson')
cor_target = abs(cor['FINAL_LABEL'])
relevant_features = abs(cor_target) [cor_target >= 0.10]
relevant_features

#8888888 Spearman Corealtion ------------------------------------------------------
plt.figure(figsize=(6, 4))
heatmap = sns.heatmap(np.round((df_final_master_table_num.corr('spearman')),2), vmin=0, vmax=1, annot
heatmap.set_title('Correlation Heatmap - spearman', fontdict={'fontsize':14}, pad=2);
# Spearman with class variable ---------------------------------------------

df_final_master_table_num.corr(method='spearman')[['FINAL_LABEL']].sort_values(by='FINAL_LABEL', asce
plt.figure(figsize=(6, 4))
heatmap = sns.heatmap(np.round((df_final_master_table_num.corr(method='spearman')[['FINAL_LABEL']]),2
heatmap.set_title('Features Correlating with Class Label - Spearman', fontdict={'fontsize':14}, pad=2
```

D. Sample predicted 100 Applicants



ID	CODE_GEN	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INCO	NAME_IN	NAME_ED	NAME_FA	NAME_HO	DAYS_BIRT	DAYS_EMP	FLAG_MO	FLAG_WO	FLAG_PHO	FLAG_EMA	OCCUPATI	CNT_FAM_	AGE_IN_Y	EMPLOYED	PREDICTION
5105143	M	Y	Y	1	90000	Working	Secondary	Married	House / ap	14263	2553	1	1	1	0	Laborers	3	39	7	PROBABLE_GOOD_CUSTOMER
5105144	M	Y	Y	1	90000	Working	Secondary	Married	House / ap	14263	2553	1	1	1	0	Laborers	3	39	7	PROBABLE_GOOD_CUSTOMER
5512969	M	Y	Y	1	90000	Working	Secondary	Married	House / ap	14263	2553	1	1	1	0	Laborers	3	39	7	PROBABLE_GOOD_CUSTOMER
5105145	F	N	Y	0	90000	Working	Higher edu	Married	House / ap	18940	1763	1	0	0	0	Sales staff	2	52	5	PROBABLE_BAD_CUSTOMER
5105146	F	N	Y	0	90000	Working	Higher edu	Married	House / ap	18940	1763	1	0	0	0	Sales staff	2	52	5	PROBABLE_BAD_CUSTOMER
5105147	F	N	Y	0	90000	Working	Higher edu	Married	House / ap	18940	1763	1	0	0	0	Sales staff	2	52	5	PROBABLE_BAD_CUSTOMER
5105150	F	N	Y	0	90000	Working	Higher edu	Married	House / ap	18940	1763	1	0	0	0	Sales staff	2	52	5	PROBABLE_BAD_CUSTOMER
5105156	F	Y	Y	2	112500	Working	Higher edu	Married	House / ap	13405	3252	1	0	1	0	Accountan	4	37	9	PROBABLE_BAD_CUSTOMER
5105158	F	Y	Y	2	112500	Working	Higher edu	Married	House / ap	13405	3252	1	0	1	0	Accountan	4	37	9	PROBABLE_BAD_CUSTOMER
5512971	F	Y	Y	2	112500	Working	Higher edu	Married	House / ap	13405	3252	1	0	1	0	Accountan	4	37	9	PROBABLE_BAD_CUSTOMER
5512972	F	Y	Y	2	112500	Working	Higher edu	Married	House / ap	13405	3252	1	0	1	0	Accountan	4	37	9	PROBABLE_BAD_CUSTOMER
5105159	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105160	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105161	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105162	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105163	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105164	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105165	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105166	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105167	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105168	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105169	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105170	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105171	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105172	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105173	M	N	Y	2	450000	Commerci	Higher edu	Married	House / ap	12238	2187	1	1	1	0	Managers	4	34	6	PROBABLE_GOOD_CUSTOMER
5105178	M	N	Y	0	180000	Working	Secondary	Married	House / ap	16165	848	1	1	0	0	Laborers	2	44	2	PROBABLE_GOOD_CUSTOMER
5105181	M	N	Y	0	180000	Working	Secondary	Married	House / ap	16165	848	1	1	0	0	Laborers	2	44	2	PROBABLE_GOOD_CUSTOMER
5105183	M	N	Y	0	180000	Working	Secondary	Married	House / ap	16165	848	1	1	0	0	Laborers	2	44	2	PROBABLE_GOOD_CUSTOMER
5105184	M	N	Y	0	180000	Working	Secondary	Married	House / ap	16165	848	1	1	0	0	Laborers	2	44	2	PROBABLE_GOOD_CUSTOMER
5105185	M	N	Y	0	180000	Working	Secondary	Married	House / ap	16165	848	1	1	0	0	Laborers	2	44	2	PROBABLE_GOOD_CUSTOMER
5105186	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105187	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105188	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105189	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105190	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105191	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105192	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105193	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105194	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105195	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105196	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105197	M	N	Y	0	315000	Commerci	Higher edu	Single / no	House / ap	10499	545	1	0	0	1	Laborers	1	29	1	PROBABLE_GOOD_CUSTOMER
5105200	M	Y	Y	2	180000	Working	Higher edu	Married	House / ap	17330	848	1	0	0	0	Drivers	4	47	2	PROBABLE_BAD_CUSTOMER
5105205	M	Y	Y	2	180000	Working	Higher edu	Married	House / ap	17330	848	1	0	0	0	Drivers	4	47	2	PROBABLE_BAD_CUSTOMER
5105206	M	Y	Y	2	180000	Working	Higher edu	Married	House / ap	17330	848	1	0	0	0	Drivers	4	47	2	PROBABLE_BAD_CUSTOMER
5105208	F	N	N	0	157500	Commerci	Secondary	Married	House / ap	16017	1398	1	0	0	0	Laborers	2	44	4	PROBABLE_GOOD_CUSTOMER
5105210	F	N	N	0	157500	Commerci	Secondary	Married	House / ap	16017	1398	1	0	0	0	Laborers	2	44	4	PROBABLE_GOOD_CUSTOMER
5105211	F	N	N	0	157500	Commerci	Secondary	Married	House / ap	16017	1398	1	0	0	0	Laborers	2	44	4	PROBABLE_GOOD_CUSTOMER
5105212	F	N	N	0	157500	Commerci	Secondary	Married	House / ap	16017	1398	1	0	0	0	Laborers	2	44	4	PROBABLE_GOOD_CUSTOMER
5105213	F	N	N	0	157500	Commerci	Secondary	Married	House / ap	16017	1398	1	0	0	0	Laborers	2	44	4	PROBABLE_GOOD_CUSTOMER
6586899	F	Y	Y	0	279000	Working	Higher edu	Married	House / ap	9949	1229	1	0	0	1	IT staff	2	27	3	PROBABLE_BAD_CUSTOMER
6586900	F	Y	Y	0	279000	Working	Higher edu	Married	House / ap	9949	1229	1	0	0	1	IT staff	2	27	3	PROBABLE_BAD_CUSTOMER
5105217	F	Y	N	1	85500	State serva	Higher edu	Married	House / ap	10044	971	1	0	0	1	Accountan	3	28	3	PROBABLE_BAD_CUSTOMER
5105218	F	Y	N	1	85500	State serva	Higher edu	Married	House / ap	10044	971	1	0	0	1	Accountan	3	28	3	PROBABLE_BAD_CUSTOMER
5105219	M	N	Y	0	135000	Working	Secondary	Single / no	Rented apa	8755	1061	1	0	0	0	Laborers	1	24	3	PROBABLE_GOOD_CUSTOMER
5105220	M	N	Y	0	135000	Working	Secondary	Single / no	Rented apa	8755	1061	1	0	0	0	Laborers	1	24	3	PROBABLE_GOOD_CUSTOMER
5105221	F	Y	Y	1	135000	Working	Secondary	Married	House / ap	15068	2067	1	0	0	0	Sales staff	3	41	6	PROBABLE_GOOD_CUSTOMER
5105222	F	Y	Y	1	135000	Working	Secondary	Married	House / ap	15068	2067	1	0	0	0	Sales staff	3	41	6	PROBABLE_GOOD_CUSTOMER
5105224	F	Y	Y	1	135000	Working	Secondary	Married	House / ap	15068	2067	1	0	0	0	Sales staff	3	41	6	PROBABLE_GOOD_CUSTOMER
5513263	F	Y	Y	1	135000	Working	Secondary	Married	House / ap	15068	2067	1	0	0	0	Sales staff	3	41	6	PROBABLE_GOOD_CUSTOMER
5105225	F	N	Y	0	135000	Working	Secondary	Married	House / ap	17365	2128	1	0	0	0	Laborers	2	48	6	PROBABLE_GOOD_CUSTOMER
5105226	F	N	Y	0	135000	Working	Secondary	Married	House / ap	17365	2128	1	0	0	0	Laborers	2	48	6	PROBABLE_GOOD_CUSTOMER
5105227	F	N	Y	0	135000	Working	Secondary	Married	House / ap	17365	2128	1	0	0	0	Laborers	2	48	6	PROBABLE_GOOD_CUSTOMER
5513306	F	N	Y	0	135000	Working	Secondary	Married	House / ap	17365	2128	1	0	0	0	Laborers	2	48	6	PROBABLE_GOOD_CUSTOMER
5105230	F	Y	N	1	157500	Working	Higher edu	Married	House / ap	11900	2828	1	0	0	0	Laborers	3	33	8	PROBABLE_BAD_CUSTOMER
5105231	F	Y	N	1	157500	Working	Higher edu	Married	House / ap	11900	2828	1	0	0	0	Laborers	3	33	8	PROBABLE_BAD_CUSTOMER
5105237	F	N	Y	0	90000	Working	Secondary	Married	House / ap	11931	271	1	1	1	0	Sales staff	2	33	1	PROBABLE_GOOD_CUSTOMER
5105238	F	N	Y	0	90000	Working	Secondary	Married	House / ap	11931	271	1	1	1	0	Sales staff	2	33	1	PROBABLE_GOOD_CUSTOMER
5105239	M	N	Y	2	180000	Working	Secondary	Married	House / ap	12998	392	1	1	0	0	Security st	4	36	1	PROBABLE_GOOD_CUSTOMER
5105240	M	N	Y	2	180000	Working	Secondary	Married	House / ap	12998	392	1	1	0	0	Security st	4	36	1	PROBABLE_GOOD_CUSTOMER
5105241	F	N	N	0	112500	Working	Secondary	Married	House / ap	14133	1214	1	0	0	0	Medicine s	2	39	3	PROBABLE_GOOD_CUSTOMER
5105242	F	N	N	0	112500	Working	Secondary	Married	House / ap	14133	1214	1	0	0	0	Medicine s	2	39	3	PROBABLE_GOOD_CUSTOMER
5105245	F	N	N	0	207000	Working	Higher edu	Married	House / ap	14740	6723	1	0	1	0	Core staff	2	40	18	PROBABLE_BAD_CUSTOMER
5105246	F	N	N	0	207000	Working	Higher edu	Married	House / ap	14740	6723	1	0	1	0	Core staff	2	40	18	PROBABLE_BAD_CUSTOMER
5105248	F	N	N	0	207000	Working	Higher edu	Married	House / ap	14740	6723	1	0	1	0	Core staff	2	40	18	PROBABLE_BAD_CUSTOMER
5105249	F	N	N	0	207000	Working	Higher edu	Married	House / ap	14740	6723	1	0	1	0	Core staff	2	40	18	PROBABLE_BAD_CUSTOMER
5105250	F	N	N	0	207000	Working	Higher edu	Married	House / ap	14740	6723	1	0	1	0	Core staff	2	40	18	PROBABLE_BAD_CUSTOMER
5105251	F	N	N	0	207000	Working	Higher edu	Married	House / ap	14740	6723	1	0	1	0	Core staff	2	40	18	PROBABLE_BAD_CUSTOMER
5105252	F	N	N	0	207000	Working	Higher edu	Married	House / ap	14740	6723	1	0	1	0	Core staff	2	40	18	PROBABLE_BAD_CUSTOMER
5105253	M	Y	Y	0	81000	Working	Secondary	Married	House / ap	13527	1153	1	0	0	0	High skill te	2	37	3	PROBABLE_GOOD_CUSTOMER
5105254	M	Y	Y	0	81000	Working	Secondary	Married	House / ap	13527	1153	1	0	0	0	High skill te	2	37	3	PROBABLE_GOOD_CUSTOMER
5105255	M	Y	Y	0	81000	Working	Secondary	Married	House / ap	13527	1153	1	0	0	0	High skill te	2	37	3	PROBABLE_GOOD_CUSTOMER
5105256	M	Y	Y	0	81000	Working	Secondary	Married	House / ap	13527	1153	1	0	0	0	High skill te	2	37	3	PROBABLE_GOOD_CUSTOMER
5105257	M	Y	Y	0	81000	Working	Secondary	Married	House / ap	13527	1153	1	0	0	0	High skill te	2	37	3	PROBABLE_GOOD_CUSTOMER
5105258	M	Y	Y	0	81000	Working	Secondary	Married	House / ap	13527	1153	1	0	0	0	High skill te	2	37	3	PROBABLE_GOOD_CUSTOMER
5105259	M	Y	Y	0	81000	Working	Secondary	Married	House / ap	13527	1153	1	0	0	0	High skill te	2	37	3	PROBABLE_GOOD_CUSTOMER
5105260	M	N	Y	0	180000	Working	Incomplete	Single / no	House / ap	10305	210	1	0	0	1	Laborers	1	28	1	PROBABLE_GOOD_CUSTOMER
5105261	M	N	Y	0	180000	Working	Incomplete	Single / no	House / ap	10305	210	1	0	0	1	Laborers	1	28	1	PROBABLE_GOOD_CUSTOMER
5105262	M	N	Y	0	180000	Working	Incomplete	Single / no	House / ap	10305	210	1	0	0	1	Laborers	1	28	1	PROBABLE_GOOD_CUSTOMER
5105263	M	N	Y	0	180000	Working	Incomplete	Single / no	House / ap	10305	210	1	0	0	1	Laborers	1	28	1	PROBABLE_GOOD_CUSTOMER
5105264	F	N	Y	2	81000	Working	Secondary	Married	House / ap	10705	149	1	0	0	0	Core staff	4	29	0	PROBABLE_GOOD_CUSTOMER
5105265	F	N	Y	2	81000	Working	Secondary	Married	House / ap	10705	149	1	0	0	0	Core staff	4	29	0	PROBABLE_GOOD_CUSTOMER
5105266	F	N	Y	2	81000	Working	Secondary	Married	House / ap	10705	149	1	0	0	0	Core staff	4	29	0	PROBABLE_GOOD_CUSTOMER
5513477	F	N	Y	2	81000	Working	Secondary	Married	House / ap	10705	149	1	0	0	0	Core staff	4	29	0	PROBABLE_GOOD_CUSTOMER
5105280	F	N	Y	1	90000	Working	Secondary	Civil marria	House / ap	18682	514	1	0	1	0	Laborers	3	51	1	PROBABLE_GOOD_CUSTOMER
5105283	M	Y	N	0	225000	Working	Secondary	Single / no	House / ap	10866	636	1	0	0	0	Drivers	1	30	2	PROBABLE_GOOD_CUSTOMER
5105288	M	Y	N	0	225000	Working	Secondary	Single / no	House / ap	10866	636	1	0	0	0	Drivers	1	30	2	PROBABLE_GOOD_CUSTOMER
5105289	M	Y	N	0	225000	Working	Secondary	Single / no	House / ap	10866	636	1	0	0	0	Drivers	1	30	2	PROBABLE_GOOD_CUSTOMER
5105290	F	N	N	0	112500	Working	Secondary	Widow	With paren	10240	818	1	1	1	0	Core staff	1	28	2	PROBABLE_BAD_CUSTOMER

REFERENCES

Agarwal, A., Rana, A., Gupta, K., Verma, N., 2020. A Comparative Study and enhancement of classification techniques using Principal Component Analysis for credit card dataset, in: 2020 International Conference on Intelligent Engineering and Management (ICIEM). Presented at the 2020 International Conference on Intelligent Engineering and Management (ICIEM), IEEE, London, United Kingdom, pp. 443–448. https://doi.org/10.1109/ICIEM48762.2020.9160230

Antonakis, A.C., Sfakianakis, M.E., 2009. Assessing naïve Bayes as a method for screening credit applicants. Journal of Applied Statistics 36, 537–545. https://doi.org/10.1080/02664760802554263

Banasik, J., Crook, J., Thomas, L., 1999. Not if but when will borrowers default. Journal of the Operational Research Society 50, 6.

Bhatore, S., Mohan, L., Reddy, Y.R., 2020. Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial TechnologyL 4, 111–138. https://doi.org/10.1007/s42786-020-00020-3

Blagus, R., Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106. https://doi.org/10.1186/1471-2105-14-106

Chornous, G., Nikolskyi, I., 2018. Business-Oriented Feature Selection for Hybrid Classification Model of Credit Scoring, in: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). Presented at the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), IEEE, Lviv, pp. 397–401. https://doi.org/10.1109/DSMP.2018.8478534

Data Science Process Alliance, n.d. What is CRISP DM? What is CRISP DM? URL https://www.datascience-pm.com/crisp-dm-2/ (accessed 11.20.20).

Elreedy, D., Atiya, A.F., 2019. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences 505, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Fernando, J., 2021. Delinquent Account Credit Card [WWW Document]. Loan Basis. URL https://www.investopedia.com/terms/d/delinquent-account-credit-card.asp (accessed 4.28.21).

Galarnyk, M., 2018. Understanding Boxplots. Understanding Boxplots. URL https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 (accessed 1.20.21).

Hussein, A.S., Li, T., Yohannese, C.W., Bashir, K., 2019. A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE. International Journal of Computational Intelligence Systems 12(2), 11.

Jafar Hamid, A., Ahmed, T.M., 2016. Developing Prediction Model of Loan Risk in Banks Using Data Mining. MLAIJ 3, 1–9. https://doi.org/10.5121/mlaij.2016.3101

Karthiban, R., Ambika, M., Kannammal, K.E., 2019. A Review on Machine Learning Classification Technique for Bank Loan Approval, in: 2019 International Conference on Computer Communication and Informatics (ICCCI). Presented at the 2019 International Conference on Computer Communication and Informatics (ICCCI), IEEE, Coimbatore, Tamil Nadu, India, pp. 1–6. https://doi.org/10.1109/ICCCI.2019.8822014

Kumar, B., 2021. 10 Techniques to deal with Imbalanced Classes in Machine Learning. 10 Techniques to deal with Imbalanced Classes in Machine Learning. URL https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

Kumar Gupta, D., Goyal, S., 2018. Credit Risk Prediction Using Artificial Neural Network Algorithm. IJMECS 10, 9–16. https://doi.org/10.5815/ijmecs.2018.05.02

Lee, T., Chen, I., 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications 28, 743–752. https://doi.org/10.1016/j.eswa.2004.12.031

Leo, M., Sharma, S., Maddulety, K., 2019. Machine Learning in Banking Risk Management: A Literature Review. Risks 7, 29. https://doi.org/10.3390/risks7010029

Madyatmadja, E.D., Aryuni, M., 2005. Comparative Study of Data Mining Model for Credit Card Application Scoring in Bank. Journal of Theoretical and Applied Information Technology 59, 6.

Marqués, A.I., García, V., Sánchez, J.S., 2012. Two-level classifier ensembles for credit risk assessment. Expert Systems with Applications 39, 10916–10922. https://doi.org/10.1016/j.eswa.2012.03.033

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J., Ryu, K., 2019. An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. Sustainability 11, 699. https://doi.org/10.3390/su11030699

Nellur, S., 2020. The Neural Network at its Simplest. The Neural Network at its Simplest. URL https://medium.com/analytics-vidhya/neural-networks-in-nutshell-7d1cc3ae6443 (accessed 2.25.21).

Oreski, S., Oreski, D., Oreski, G., 2012. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Systems with Applications 39, 12605–12617. https://doi.org/10.1016/j.eswa.2012.05.023

Payments and Settlements Department, 2020. Payments Bulletin - Thrird Quarter (Quarterly Report), Payment Bulletin. Central Bank of Sri Lanka.

Pristyanto, Y., Adi, S., Sunyoto, A., 2019. The Effect of Feature Selection on Classification Algorithms in Credit Approval, in: 2019 International Conference on Information and Communications Technology (ICOIACT). Presented at the 2019 International Conference on Information and Communications Technology (ICOIACT), IEEE, Yogyakarta, Indonesia, pp. 451–456. https://doi.org/10.1109/ICOIACT46704.2019.8938523

Ray, S., n.d. Understanding Support Vector Machine(SVM) algorithm from examples (along with code). URL https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (accessed 3.25.21).

Sariannidis, N., Papadakis, S., Garefalakis, A., Lemonakis, C., Kyriaki-Argyro, T., 2019. Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques. Annals of Operations Research. https://doi.org/10.1007/s10479-019-03188-0

Saunders, A., Cornett, M.M., 2014. Financial Institutions Management: A risk Management Approach, Eighth edition. ed. McGraw-Hill Education, New York, NY.

Sharma, S., 2017. Activation Functions in Neural Networks. Sigmoid, tanh, Softmax, ReLU, Leaky ReLU EXPLAINED !!! URL https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

Shoumo, S.Z.H., Dhruba, M.I.M., Hossain, S., Ghani, N.H., Arif, H., Islam, S., 2019. Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking, in: TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON). Presented at the TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), IEEE, Kochi, India, pp. 2023–2028. https://doi.org/10.1109/TENCON.2019.8929527

Song, X., 2019. A Credit Card Dataset for Machine Learning. Credit Card Approval Prediction. URL https://www.kaggle.com/rikdifos/credit-card-approval-prediction (accessed 9.10.20).

SriLaxmi, K., Divya, N, Lakshmi, P., Vidya, A., Hameeda, S., 2020. Credit Card Customer Predicting using Machine Learning. International Journal for Research in Applied Science and Engineering Technology 8, 2697–2701. https://doi.org/10.22214/ijraset.2020.5452

Statistics Solutions, n.d. Correlation (Pearson, Kendall, Spearman). Correlation. URL https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/ (accessed 2.7.21).

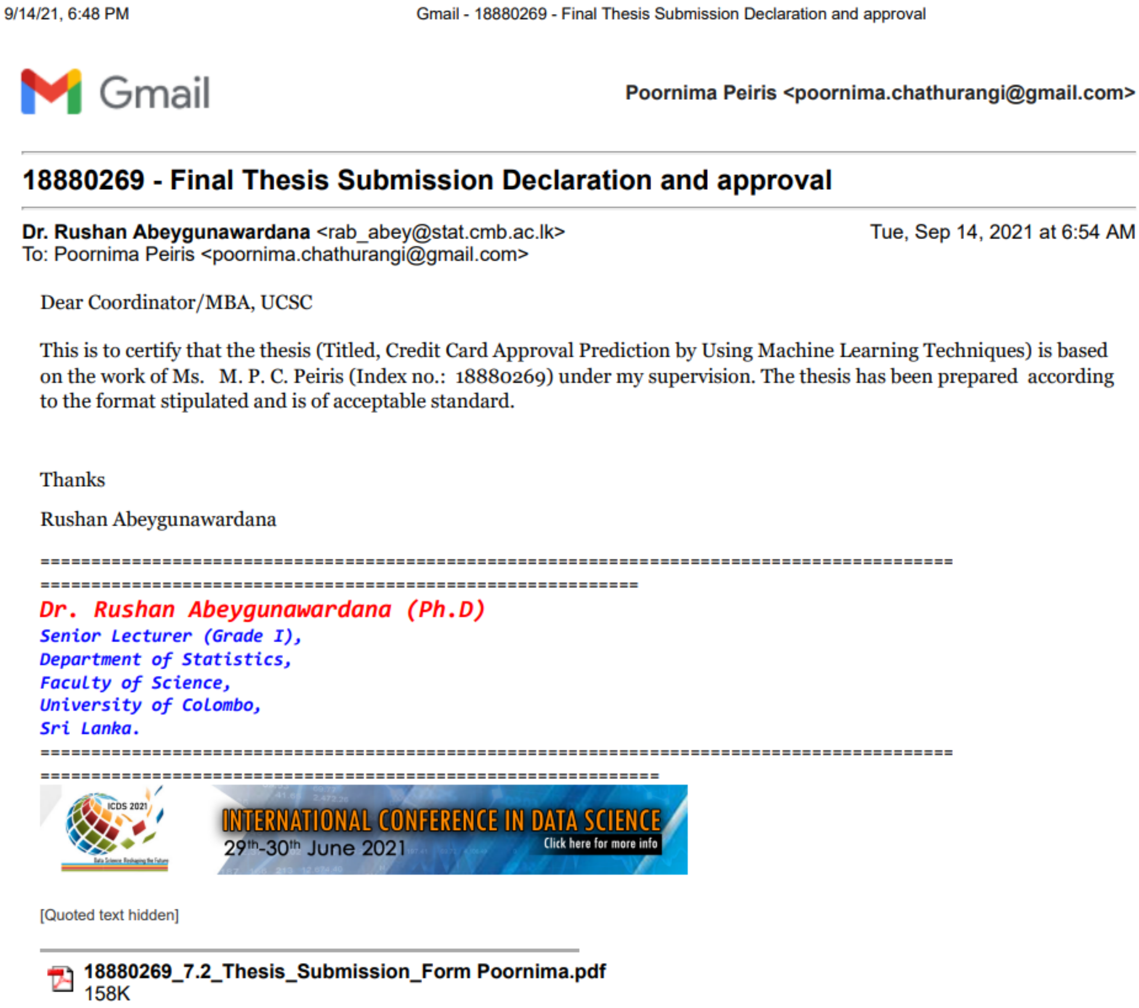
Taylor, J., 2017. Four Problems in Using CRISP-DM and How To Fix Them. Decision Management Solutions. URL https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html (accessed 12.20.20).

Thomas J. Catalano, A.B., 2020. What Is a Credit Card? Credit Cards. URL https://www.investopedia.com/terms/c/creditcard.asp (accessed 12.10.20).

Wang, G., Hao, J., Ma, J., Jiang, H., 2011. A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications 38, 223–230. https://doi.org/10.1016/j.eswa.2010.06.048

Wang, Y., Zhang, Y., Lu, Y., Yu, X., 2020. A Comparative Assessment of Credit Risk Model Based on Machine Learning a case study of bank loan data. Procedia Computer Science 174, 141–149. https://doi.org/10.1016/j.procs.2020.06.069

Supervisors' Approval



9/14/21, 6:48 PM Gmail - 18880269 - Final Thesis Submission Declaration and approval

Gmail Poornima Peiris <poornima.chathurangi@gmail.com>

18880269 - Final Thesis Submission Declaration and approval

Dr. Rushan Abeygunawardana <rab_abey@stat.cmb.ac.lk> Tue, Sep 14, 2021 at 6:54 AM
To: Poornima Peiris <poornima.chathurangi@gmail.com>

Dear Coordinator/MBA, UCSC

This is to certify that the thesis (Titled, Credit Card Approval Prediction by Using Machine Learning Techniques) is based on the work of Ms. M. P. C. Peiris (Index no.: 18880269) under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Thanks

Rushan Abeygunawardana

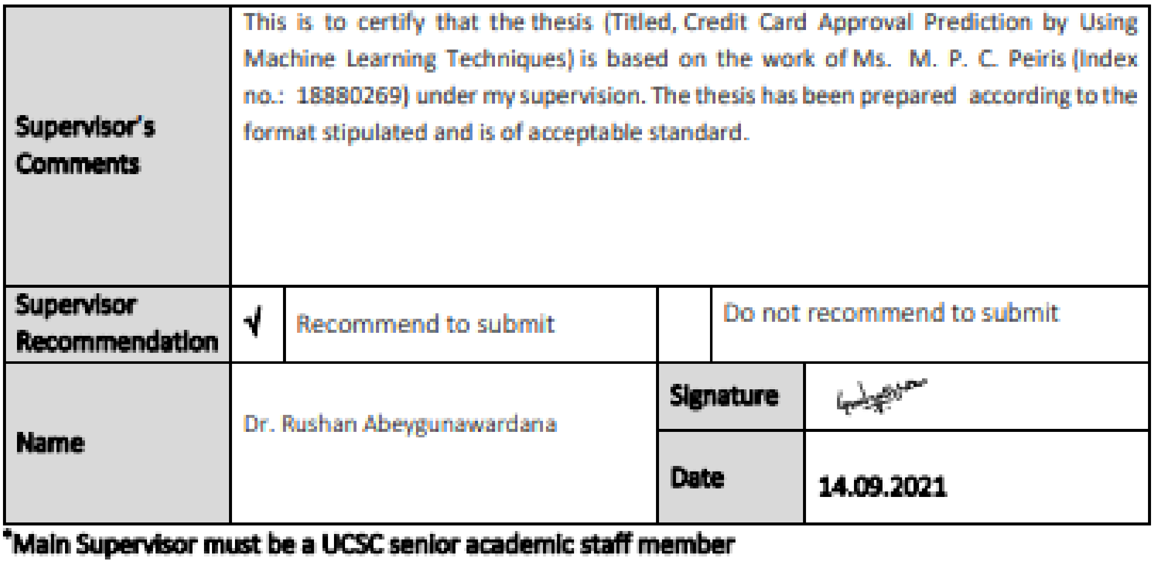
==
===

Dr. Rushan Abeygunawardana (Ph.D)
Senior Lecturer (Grade I),
Department of Statistics,
Faculty of Science,
University of Colombo,
Sri Lanka.

==
===

[Quoted text hidden]

18880269_7.2_Thesis_Submission_Form Poornima.pdf
158K



Supervisor's Comments	This is to certify that the thesis (Titled, Credit Card Approval Prediction by Using Machine Learning Techniques) is based on the work of Ms. M. P. C. Peiris (Index no.: 18880269) under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.			
Supervisor Recommendation	√	Recommend to submit		Do not recommend to submit
Name	Dr. Rushan Abeygunawardana	**Signature**		
		Date	**14.09.2021**	

***Main Supervisor must be a UCSC senior academic staff member**