A PROJECT REPORT ON

Default of Credit Card Clients

A Project report submitted in

Partial fulfillment of the requirement for the award of the Degree of

MASTER

IN

COMPUTER APPLICATION

Submitted By

Animesh Panda

(Regd. No-2224100004)

Under the esteemed guidance of

Dr. Debasis Gountia

Associate Professor

School of Computer Sciences

School of Computer Sciences

ODISHA UNIVERSITY OF TECHNOLOGY & RESEARCH

(Formerly College of Engineering & Technology), ODISHA

Techno Campus, Ghatikia, Mahalaxmi Vihar, Bhubaneswar-751029

Year 2023-2024

ODISHA UNIVERSITY OF TECHNOLOGY & RESEARCH

(Formerly College of Engineering & Technology), ODISHA

Techno Campus, Ghatikia, Mahalaxmi Vihar, Bhubaneswar-751029

School of Computer Sciences

CERTIFICATE

This is to certify that the project report entitled Default of Credit Card Clients being submitted by Mr Animesh Panda in partial fulfillment for the award of the Degree of Master in Computer Application to the Odisha University of Technology & Research is a record of bonafied work carried out by him under my guidance and supervision.

The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree or Diploma.

Signature of Internal Guide

Name:

Designation:

Signature of
Head of School of Computer Sciences

Name:

Designation:

ODISHA UNIVERSITY OF TECHNOLOGY & RESEARCH

(Formerly College of Engineering & Technology), ODISHA

Techno Campus, Ghatikia, Mahalaxmi Vihar, Bhubaneswar-751029

School of Computer Sciences

DECLARATION

I **Animesh Panda** bearing Registration No: **2224100004,** a bonafide student of **Odisha University of Technology & Research,** would like to declare that the project titled **"Default of Credit Card Clients"** A partial fulfillment of MCA Degree course of Odisha University of Technology & Research is my original work in the year 2023 under the guidance of **Dr. Debasis Gountia,** Associate Professor, School of Computer Sciences and it has not previously formed the basis for any degree or diploma or other any similar title submitted to any university.

Animesh Panda

Date: **Regd. No-2224100004**

ODISHA UNIVERSITY OF TECHNOLOGY & RESEARCH

(Formerly College of Engineering & Technology), ODISHA

Techno Campus, Ghatikia, Mahalaxmi Vihar, Bhubaneswar-751029

School of Computer Sciences

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my advisor, **Dr. Debasis Gountia,** Associate Professor, School of Computer Sciences, whose knowledge and guidance has motivated me to achieve goals I never thought possible. He has consistently been a source of motivation, encouragement, and inspiration. The time I have spent working under his supervision has truly been a pleasure.

I take it as a great privilege to express our heartfelt gratitude to **Prof. Dr. Jibitesh Mishra**, Head of School of Computer Sciences for his valuable support and all senior faculty members of School of Computer Science for their help during my course. Thanks to programmers and non-teaching staff of School of Computer Science of OUTR.

I would also like to extend my gratitude to respected **Hon'ble Vice Chancellor Prof. Dr. Bibhuti Bhusan Biswal** for extending his utmost support and co-operation in providing all the provisions, and Management for providing excellent facilities to carry out my project work.

Finally special thanks to my parents, sister for their support and encouragement throughout my life and this course. Thanks to all my friends and well-wishers for their constant support.

Animesh Panda

Regd. No-2224100004

ABSTRACT

In modern day's credit card plays an important role in every person's daily activity. Customer purchases their needs with their credit cards and online transitions. Banks and financial institutes consider denying the credit applications of customers to avoid the risk of defaulters. Credit risk is the rise of debt on the customer who fails to make the billing payment for some period. The purpose of the project is how to reduce the defaulters among the list of customers, and make a background check on whether to provide the loan or not and to find the promising customers. These predictive models would benefit the lending institutions and to the customers as it would make them more aware of their potential defaulting rate. The problem is a binary classification problem whether a customer will be defaulting to pay next month payment. The dataset is unbalanced so the focus was on the precision and recall more than the accuracy metrics. After comparison with precision-recall curve, logistic regression is the best model based on the False Negative value of confusion metrics. Moreover, after changing the threshold value of the logistic regression, GUI (Graphical user interface) implemented and predicted whether a customer is defaulter or not-defaulter.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Background and Information of the Study

Credit card is a physical card used for paying our bills easily. The cardholder could use it to give a paying promise as a requital to the cost of service and goods. There is a brief explanation of algorithms to define term credit scoring [1], which determines the relation between defaulters and loan characteristics. It is a useful information for financial institution to maintain financial statement and customer transaction list to reduce the uncertainty. Yeh and Lien (2009) compared the predictive accuracy of probability of default among six data mining methods (specifically, K-nearest neighbor classifier, logistic regression, discriminant analysis, naive Bayesian classifier, artificial neural networks, and classification trees) using customers default payments data in Taiwan. Their experimental results indicated that only artificial neural network could accurately estimate default probability. The use of Taiwan data is beneficial for us because the sample size of the default payment data in Taiwan is 30,000. [2]

Currently, a variety of Machine Learning approaches used to detect fraud and predict payment defaults. Some of the more common techniques include Logistic Regression, K Nearest Neighbor, Decision Tree, Naive Bayes, Support Vector Machine, Feed Forward Neural Networks and Ensemble approaches like Voting Classifier. The dataset contains information on 24 variables, obtained from the UCI Machine Learning Repository. Here we categorized the dataset based on independent variables such as credit amount, age, sex, education, marital status, and their past loan repayment history of last 6 months, History of their past payments made (April to September), amount of bill statement, amount of previous payment. The dependent variable is default, which means whether the customer will pay their next month payment, or not. We can reduce the cost, make a good decision for a potential customer and help in reducing the time consumption for processing loan application and more.

1.2 Statement of Problem

Credit card is a physical card used for paying our bills easily. The cardholder could use it to give a paying promise as a requital to the cost of service and goods. There is a brief explanation of algorithms to define term credit scoring [1], which determines the relation between defaulters and loan characteristics. It is a useful information for financial institution to maintain financial statement and customer transaction list to reduce the uncertainty. Yeh and Lien (2009) compared the predictive accuracy of probability of default among six data mining methods (specifically, K-nearest neighbor classifier, logistic regression, discriminant analysis, naive Bayesian classifier, artificial neural networks, and classification trees) using customers default payments data in Taiwan. Their experimental results indicated that only artificial neural network could accurately estimate default probability. The use of Taiwan data is beneficial for us because the sample size of the default payment data in Taiwan is 30,000.

1.3 Aim and Objective of the Study

The problem is to classify the defaulters and non-defaulters on the credit payment of the customers. This project is helpful for solving the real problem by using various classification techniques. Moreover, any user can access GUI and add their gender, education, marital status and payment details to check next month in which category they fall (defaulter or non-defaulter).

The core objectives: Find whether the customer could pay back his next credit amount or not and Identify some potential customers for the bank who can settle their credit balance. The steps followed to manage these goals:

- Selection of dataset
- Display some graphical information and visualize the features.
- Check Null values in the dataset
- Data pre-processing using one-hot encoding and remove extra parameters
- Train with classifiers
- Evaluate the model with test data
- Compare the accuracy, precision and recall finding the optimal model.

• Created a Graphical User Interface to check with real time customer data and predict defaulter for their next month payment

B. High-level overview

The major purpose of risk prediction is to use information, such as financial statement, customer transaction and repayment records to predict individual customer's credit risk and to reduce the damage and uncertainty. Many methods, including Logistic Regression, SVM, KNN, Decision Tree, Naive Bayes and Feed Forward Artificial Neural Networks used to develop models of risk prediction. [4]

The remainder of this paper organized as follows. Section 2 summarizes the basic properties of applied models and accuracy explores the methodology with data preprocessing. Section 4 comprises of evaluation process. Section 5 presents summary.

CHAPTER TWO

SOFTWARE AND HARDWARE REQUIREMENTS

Computer system is made up of units that are put together to work as one in order to achieve a common goal. The requirements for the implementation of the new system are:

- The Hardware
- The Software

A. Hardware Requirements

These are the physical component needed by the system to operate.

- 500mb of Ram(Minimum)
- Keyboard
- Mouse
- Printer
- Intel Pentium

B. Software Requirements

- Processor speed- 1.30Hz and above
- Web browser
- Operating system
- Coding Language : Python 3.8
- Web Framework : Flask

CHAPTER THREE

LITERATURE REVIEW

3.1 Background of the Study

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

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 fulfil 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 analysing 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.

3.2 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 reach 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).

3.3 Objectives of the Study

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

3.4 Research Question

There are many classification learning algorithms that used to predict and select the best customer for approving the Credit Cards and this will result to decrease the non

performing Credit Card percentage of the banking system. Further, it is important to study whether the different algorithms behave differently. Therefore, following research question need to be addressed.

3.5 Scope of the Study

According to the data in relating to non-performing loans and advances in relating to loans and advances of the banking sector of Sri Lanka drastically increasing during past decade. Non-Performing loans and advances ratios of banks increasing over 5% and it reduce the profitability of the banks and increase the risks towards the banking sector. Especially in defaults in credit cards are increasing drastically in recent years. Customers are become defaulters of credit cards willingly. This trend affects negatively to the banking sector. Managers and providers of credits cards in banking and finance sector must have ability to identify the credit card defaulters easily. But various Banks use various kind of credit scoring models and risk analysing models using both statistical and machine learning approaches. In most Sri Lankan banks use manual methods to identify the credit risks of consumers.

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

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

3.5 Feasibility of the Study

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

Machine learning models are used rarely to identify credit card defaulters. Bank of Ceylon has issued over 100,000 credit cards to their customers and maintain database of the credit cards. This database includes both performing and non-performing credit cards. Management of the Bank provide access to use this database to predict credit card defaults

using machine learning approaches. Therefore, this study is feasible and beneficial to the researcher as well as Bank.

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

3.6 Supervised and Unsupervised Learning

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

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

3.7 Default prediction models

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

There is much research on credit card lending. It is a widely researched subject. Many statistical methods have applied to developing credit risk prediction, such as Logistic Regression, Support Vector machine, K-nearest neighbour classifiers, probabilistic classifiers such as Bayes classifiers and neural networks and ensemble classifiers such as Voting Classifier.

A. K-nearest neighbour:

K-nearest neighbour (KNN) is one of the simplest supervised classifiers. The vision is to define K centroids, one for each cluster. These centroids placed inappropriately because of different location causing different results. Selecting the value of K is more critical

part as a small value of K means that noise will have a higher influence on the result (probability of over fitting is high) and on another side, the higher value of K defeat idea to find the nearest value and lead to a greater amount of time & under fitting of model. When given an unknown data, the KNN classifier searches the pattern space for the KNN, which are the closest to this unknown data.

B. Gaussian naïve Bayes (GNB)

The Bayesian classifier is a probabilistic classifier based on Bayes theorem. Naïve Bayes classifier assumes that all features are unrelated to each other and more useful for predictive modelling. In practice, however, dependences can exist between variables. Advantages of naïve bayes are easy and fast to predict the class of test dataset. It performs well in multi-class prediction. Limitation of naïve bayes is the assumption of independent predictors, is hard to get a set of predictors, which are completely independent. [5]

C. Logistic Regression

Logistic Regression is a form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictors. Binary logistic regression used when the dependent variable is binary or has two levels. In logistic regression, the purpose of the analysis is to assess the effects of multiple explanatory variables, which can be numeric or categorical. The goal is to find the best fitting model to describe the relationship between the binary characteristic of interest and a set of independent variables. [5] D. Decision tree: A decision tree is a flowchart-like structure in which each inward node represents a "test" on an attribute, each branch represents the consequence of the test, and each leaf node represents a class label. In decision analysis, a decision tree and the closely related influence diagram used as a visual and analytical decision support tool, where the expected values of competing alternatives are calculated.

E. Support Vector Machine

Support vector machine is a popular machine learning classification algorithm. SVM used as supervised learning when the dataset has features and class labels. The Linear classifier implemented in a code. A focus is to maximize the distance from hyper plane to the nearest data point of either class in SVM; the maximum-margin hyper plane determined by the dataset lies nearest to it. These data points which influences hyper plane knows as Supper vector. When data separated linearly, draw two parallel hyper planes, which

separate two classes of data. The Distance between two hyper planes is 2/||w||, to maximize this distance denominator value should be minimized i.e., ||w|| should be minimized. (Shows in image 1.1). The different kernel is available as linear, poly, sigmoid and wrong choice of the kernel can lead to an increase in error percentage. [6]

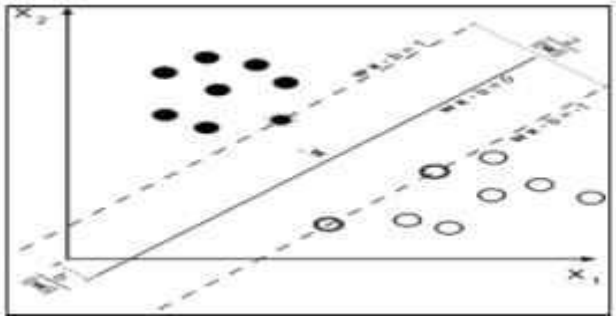


Fig (1.1) Support Vector Machine with hyper plane and two classes

F. Feed Forward Artificial Neural Networks

A feed forward neural network is a type of artificial neural network. It can perform several classification tasks at once, although commonly each network performs only one. The best result is usually to train separate networks for each output, then to combine them into an ensemble so that they can run as a unit. In Feed forward neural network, all nodes connected in a network. Predictions based on the input nodes and weights. As the name suggests, activation flow is from the input layer to the output layer. There is one or more hidden layer between the input and output layer and no cycle or loop into the network. One neuron called a perceptron. A perceptron consists of one input layer and one neuron. A total node in the input layer is the same as total features in the dataset. Each input multiplied with random weight value where a weight is in the range of 0 to 1. An activation function knows as summed weighted input to the output of neurons. Neurons have activation function such as a step function, sigmoid, relu, softmax or tanh function. Bias added to the sum of input and weight to avoid null values. [7]

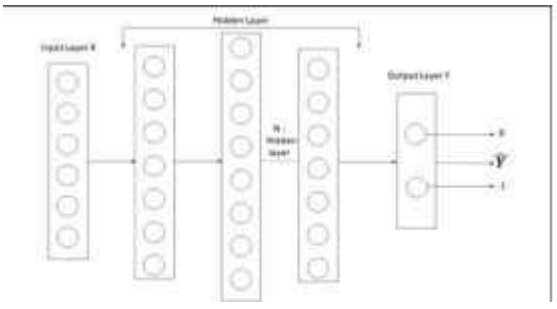


Fig (1.2): Neural Network of this dataset

G. Ensemble Learning

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Voting Classifier is one of the simplest ways of combining the predictions from multiple machine learning algorithms by first creating two or more standalone models from your training dataset. A Voting Classifier used to wrap your models and average the predictions of the sub-models when asked to make predictions for new data. The predictions of the sub-models weighted, but specifying the weights for classifiers manually or even heuristically is difficult.

H. Accuracy Measures:

a. Accuracy:

The accuracy of a model is usually determined after the model parameters learned and fixed and no learning is taking place. Then the test samples fed to the model and the number of mistakes (zero-one loss) the model makes recorded, after comparison to the true targets. Then the percentage of misclassification is calculated. In our dataset, accuracy determine how often the model predicts defaulters and non-defaulters correctly.

b. Precision:

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. If the precision is high, then there will be low false positive rate. Here precision tells us that whenever our models predicts it is defaulter how often it is correct.

c. Recall:

Recall is the ratio of correctly predicted positive observations to the all observations in actual class. In other words, out of all positive class how much we have predicted correctly. When we apply this in our dataset, it shows the actual defaulters that the model will actually predict.

d. Precision Recall Curve:

It will measure the success of prediction, when classes are imbalanced. It will show the trade-off between precision and recall threshold. [8]

Table (1.1) Precision Recall Curve

#	Non-defaulter (predicted) - 0	Defaulter (predicted) - 1
Non-defaulter (actual) - 0	TN	FP
Defaulter (actual) - 1	**FN**	TP

Loss: Loss functions let the optimization function know how well it is doing. Loss functions used in the output layer, Layers that support unsupervised layer wise pre-training.

a. Cross Entropy loss: Cross-entropy loss, or log loss, measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label.

b. Binary Cross Entropy: In binary classification, where the number of classes equals 2 either 0 or 1, then it is known as binary cross entropy. [9]

Binary cross-entropy calculated as:

$$-(\log(p)+(1-y)\log(1-p)$$

CHAPTER FOUR

SOFTWARE REQUIREMENT ANALYSIS

Classical Machine Learning Techniques used for credit default prediction When considering classical machine learning techniques Neural Networks, Random Forest, K Nearest Neighbor, Support Vector Machines and Decision Trees are the successful. techniques that given better results (M.Jayadev, N.M. Shah, R. Vadlamani, 2019).

Decision Trees

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

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

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

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

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

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

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

K – Nearest Neighbour

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

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

Artificial Neural Networks (ANN)

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

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

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

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

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

Support Vector Machine (SVM)

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

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

Ensemble classifiers used for credit default prediction

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

Random Forest Algorithm

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

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

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

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

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

Islam et al (S.R. Islam, W. Eberle, S.K. Ghafoor, 2019) predict the default credit card using combine approaches of machine learning and they have identified that Random Forest perform well by showing accuracy of 94.46%, precision of 94.78%, Recall of 79.32% and F-Score of 0.8637. when comparing to other machine learning and heuristic approaches.

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

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

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

Attributes Used in Credit Default Prediction

Researches used various attributes to develop machine learning models in relating to Credit card default prediction.
Goyal and Kaur (A. Goyal, R. Kaur, 2016) used Loan ID, Gender, Marital Status, Number of Dependents, Education Level, Employment details, Income level, Loan amount, Credit History, Property area and Loan Status as main attributes.
Neema and Soibam (S. Neema and B. Soibam, 2017) used 23 variables to predict Credit Card defaults. Credit amount, Gender, Education, Marital Status, Age, History of past Payments, Amount of Bill statement, Amount of Previous payments taken as main attributes and sub attributes including under the main attributes.
Hamid and Ahmed (J.A Hamid and M.T. Ahmed, 2016) used seven attributes to predict risk of loans. Credit history, Purpose, Gender, Credit amount, Age, Housing, Job and Class is the seven attributes that taken into consideration.
Gultekin and Sakar (B. Gultekin and E.B. Sakar, 2018) used eighteen attributes to predict default credits. Housing maturity, marital status, occupation, educational status, vehicle maturity, consumer maturity, productNum, working time, workplace, ownership code, age, insurance, class, Loan type, Credit Reporting Agency and default number are used as attributes.
Supriya et al (P. Supriya, M. Pavani, N. Saisushma, V.N. Kumari, K. Vikas, 2019) used 12 attributes for loan prediction. Gender, Marital Status, Dependents, education, employment, income, amount of loan, credit history taken as major attributes.
Tejaswini et al (J. Tejaswini, M.T. Kavya, N.D.R. Ramya, S.P. Triveni, R.V. Maddumala, 2020) used 13 attributes to develop accurate credit prediction model using machine learning. Loan ID, Gender, marital status, dependents, education, employment, income credit history, property area, loans status used as main attributes.
Most of studies has divided the data set as 70% of training data and 30% of test data. Some studies divided the data set as 60% training data and 40% of test data.

Research Gap and Conclusion

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

CHAPTER FIVE
SOFTWARE DESIGN

5.1 Scope of the Study

According to the data in relating to non-performing loans and advances in relating to loans and advances of the banking sector of Sri Lanka drastically increasing during past decade. Non-Performing loans and advances ratios of banks increasing over 5% and it reduce the profitability of the banks and increase the risks towards the banking sector. Especially in defaults in credit cards are increasing drastically in recent years. Customers are become defaulters of credit cards willingly. This trend affects negatively to the banking sector. Managers and providers of credits cards in banking and finance sector must have ability to identify the credit card defaulters easily. But various Banks use various kind of credit scoring models and risk analyzing models using both statistical and machine learning approaches. In most Sri Lankan banks use manual methods to identify the credit risks of consumers.

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

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

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

5.2 Feasibility of the Study

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

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

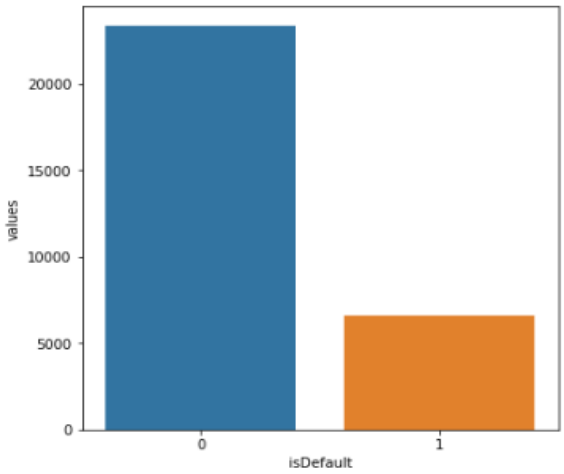
Finding of the Study will help to understand possibilities of transferring to non-performing section to Credit Cards at the initial stage of the Credit Card application process. In addition, it will help to monitor and identify the customers that can pay the Credit Card regularly and the customers that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

CHAPTER SIX

METHEDOLOGY

METHODOLOGY

In methodology, data description, independent variable and dependent variable described with scale of variables. Moreover, in the process data preprocessing and feature engineering described as below.



A. Data Description:

This dataset consists of 30000 total instances and 25 features including-

The total number of customer based on defaulter and non-defaulter from a dataset.

Figure 1.3 No: of defaulters and non-defaulters

B. Process

The first step is data preprocessing. Data preprocessing used to convert the raw data into a clean data set.

• ID column dropped as its unnecessary for our modeling.
• The attribute name 'PAY_0' converted to 'PAY_1' for naming convenience.
• Numeric attributes converted to nominal.
• One hot encoding which is a process by which categorical variables converted into a dummy form that provided to algorithms to do a better job in prediction. One hot encoder used to perform linearization of data. For instance, value in the 'EDUCATION' variables were grouped such that the values '0, 4, 5, 6' was combined to one value and assigned a value '4'.

Converting categorical features into (n-1) features. Customer ID 1 has education value 3 which is converted to 0,0,1 as 1 is assigned to high school. Likewise, for gender Male and Female respectively 0 and 1. For Marital status, there are 4 categorical values as 1 means married, 2 means single and 3 means others. As in the dataset, there is no description about value 0, so we converted to value 3 as others. So, One-hot encoding is applied to education, gender and marital status.

Robust Scaler is used which converts all the variables in the same scale so if the data contains many outliers, scaling using the mean and variance of the data is likely to not work very well then in such cases Robust Scaler is used. For example, in Limit Balance column there are different range of values, which are converted, in proper scale.

• For all classification tasks, target variable converted to numeric.

• Next step is data preparation or feature selection where features selected by declaring the independent and target variable. Different graphs like count plots and pair plots are plotted with the reference to the target variable to check the default (=0) and non-default (=1).
• Before applying algorithms on train data, dataset is split into a ratio of 60:40, which is 60% train data and 40% is test data
• Next step is to train data by applying different algorithms as Support Vector Machine, K-Neighbors Classifier, logistic regression, Gaussian Naïve Bayes and artificial neural network. Cross-validation: Cross-Validation used to assess the predictive performance of the models and to judge how they perform outside the sample to a new dataset also known as test data the reason to use cross-validation techniques is that when we fit a model, we are fitting it to a training dataset. Without cross-validation, we only have information on how our model performs in-sample data. Ideally, we would like to see how the model performs when we have new data of customers. [10]
In cross-validation process, K-fold cross validation is used. In K-fold cross validation all observations are used for both training and validation process. Normally 10-fold cross validations process is used. (Step 10). The general process of K-fold validations is to Shuffle the dataset randomly and Split the dataset into k groups (k=10)
For Neural Network, the following are tuning parameters:
Epochs: One epoch is when an entire dataset passed forward and backward via NN once. Here epoch value is set to 100.
Activation function:
ReLU: ReLU is commonly used activation function for deep learning. This has value range from zero to infinity.

$$(x)=\max(0,x)$$

Sigmoid: A sigmoid function is a differentiable, real function that defined for all real input values and has a non-negative derivative at each point.

$$(x)=1/(1+e-x)$$

SGD: Stochastic gradient descent is an iterative method for optimizing a differentiable objective function. Adam optimizer used in this project.
Input layer: Input layer is the very beginning of the workflow for neural network. 26 neurons used in input layer.
Hidden Layer: Hidden layer is in between Input Layer and Output Layer. 2 hidden layers are used after applying 1,2,3 hidden layer and found over fitting issue as we increased hidden layers.
Output Layer: It is a predicted feature value or output variables. It is an outcome. In this dataset, there are 2 neurons in an output layer.

CHAPTER SEVEN

CODE

Main.py

```
# Importing Libraries
from sklearn.ensemble import VotingClassifier
from sklearn import model_selection
from sklearn.preprocessing import RobustScaler
import matplotlib.pyplot as plt
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import LabelEncoder
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.metrics import classification_report
from sklearn import tree
import graphviz
from sklearn.utils import class_weight
print("Modules Imported! \n")
# Data preprocessing
print("--- Loading Dataset ---")
url = 'C:/Users/Kingslayer/Desktop/Project/dataset.csv'
dataset = pd.read_csv(url)
# Feature Engineering
dataset.rename(columns=lambda X: X.lower(), inplace=True)
dataset.drop('id', axis=1, inplace=True)
dataset.rename(
   columns={'default.payment.next.month': 'isDefault'}, inplace=True)
print("Dataset Info")
print("Default Credit Card Clients data -  rows:",
     dataset.shape[0], " columns:", dataset.shape[1])
dataset.describe()
# Feature Engineering
dataset['grad_school'] = (dataset['education'] == 1).astype('int')
dataset['university'] = (dataset['education'] == 2).astype('int')
pay_features = ['pay_0', 'pay_2', 'pay_3', 'pay_4', 'pay_5', 'pay_6']
for p in pay_features:
   dataset.loc[dataset[p] <= 0, p] = 0
target_name = 'isDefault'
X = dataset.drop(target_name, axis=1)
# Robust Scaler for scaling different values into proper scale
```

```
robust_scaler = RobustScaler()
X = robust_scaler.fit_transform(X)
y = dataset[target_name]
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.40, random_state=15, stratify=y)
# Defining a confusion matrix
def CMatrix(CM, labels=['non-defaulter', 'defaulter']):
    df = pd.DataFrame(data=CM, index=labels, columns=labels)
    df.index.name = 'TRUE'
    df.columns.name = 'PREDICTION'
    df.loc['Total'] = df.sum()
    df['Total'] = df.sum(axis=1)
    return df
metrics = pd.DataFrame(index=['accuracy', 'precision', 'recall'],
                       columns=['NULL', 'LogisticReg', 'ClassTree', 'NaiveBayes', 'SVM', 'KNN', 'ANN', 'VotingClassifier'])
y_pred_test = np.repeat(y_train.value_counts().idxmax(), y_test.size)
metrics.loc['accuracy', 'NULL'] = accuracy_score(
    y_pred=y_pred_test, y_true=y_test)
metrics.loc['precision', 'NULL'] = precision_score(
    y_pred=y_pred_test, y_true=y_test)
metrics.loc['recall', 'NULL'] = recall_score(y_pred=y_pred_test, y_true=y_test)
CM = confusion_matrix(y_pred=y_pred_test, y_true=y_test)
CMatrix(CM)
# A. Logistic Regression
# 1. Create an instance of the estimator
logistic_regression = LogisticRegression(n_jobs=-1, random_state=15)
# 2. Use the training data to train the estimator
logistic_regression.fit(X_train, y_train)
# 3. Evaluate the model
y_pred_test = logistic_regression.predict(X_test)
metrics.loc['accuracy', 'LogisticReg'] = accuracy_score(
    y_pred=y_pred_test, y_true=y_test)
metrics.loc['precision', 'LogisticReg'] = precision_score(
    y_pred=y_pred_test, y_true=y_test)
metrics.loc['recall', 'LogisticReg'] = recall_score(
    y_pred=y_pred_test, y_true=y_test)
 # Comparison of recall, precision and accuracy using graph
fig, ax = plt.subplots(figsize=(10, 7))
metrics.plot(kind='barh', ax=ax)
ax.grid()
# Precision-Recall Curve
fig, ax = plt.subplots(figsize=(10, 7))
ax.plot(precision_lr, recall_lr, label='LogisticReg')
ax.plot(precision_kn, recall_kn, label='KNN')
ax.plot(precision_nb, recall_nb, label='NaiveBayes')
```

```
ax.set_xlabel('Precision')
ax.set_ylabel('Recall')
ax.set_title('Precision-Recall Curve')
#ax.hlines(y=0.5, xmin=0, xmax=1, color='red')
ax.legend()
ax.grid()
# Confusion matrix for modified Logistic Regression Classifier
fig, ax = plt.subplots(figsize=(8, 5))
ax.plot(thresholds_lr, precision_lr[1:], label='Precision')
ax.plot(thresholds_lr, recall_lr[1:], label='Recall')
ax.set_xlabel('Classification Threshold')
ax.set_ylabel('Precision, Recall')
ax.set_title('Logistic Regression Classifier: Precision-Recall')
ax.hlines(y=0.48, xmin=0, xmax=1, color='red')
ax.legend()
ax.grid()
```

maingui.py

```
# Importing libraries
from tkinter import *
from collections import OrderedDict
from keras.models import model_from_json
from keras.models import load_model
from sklearn.preprocessing import RobustScaler
from sklearn.linear_model import LogisticRegression
import pandas as pd
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
# Title of GUI
master = Tk()
master.title("Credit card defaulter check")
master.geometry("300x600")
master.title('Data Analytics Project')
# Input Label
Label(master, text="Limit Balance:").grid(row=0)
Label(master, text="Male:").grid(row=1)
Label(master, text="Graduate:").grid(row=2)
Label(master, text="University:").grid(row=3)
Label(master, text="High School:").grid(row=4)
Label(master, text="Married:").grid(row=5)
Label(master, text="Single:").grid(row=6)
Label(master, text="Age:").grid(row=7)
# Get Input values
e1 = Entry(master)
e1.grid(row=0, column=1)
e2 = Entry(master)
```

```
e2.grid(row=1, column=1)
e3 = Entry(master)
e3.grid(row=2, column=1)
e4 = Entry(master)
e4.grid(row=3, column=1)
e5 = Entry(master)
e5.grid(row=4, column=1)
# Load Dataset
print("--- Loading Dataset ---")
url = 'H:/Western/Fall 2018/Data Analytics/Project/Code/dataset.csv'
dataset = pd.read_csv(url)
# Feature Engineering
dataset.rename(columns=lambda X: X.lower(), inplace=True)
dataset.drop('id', axis=1, inplace=True)
dataset.rename(
    columns={'default.payment.next.month': 'isDefault'}, inplace=True)
dataset['grad_school'] = (dataset['education'] == 1).astype('int')
dataset['university'] = (dataset['education'] == 2).astype('int')
dataset['high_school'] = (dataset['education'] == 3).astype('int')
dataset.drop('education', axis=1, inplace=True)
# Call function to check result
def make_ind_prediction(data):
    a1 = float(e1.get())
    a2 = (e2.get())
    a3 = (e3.get())
    a4 = (e4.get())
    a5 = float(e5.get())
    a6 = float(e6.get())
    a7 = float(e7.get())
    a8 = float(e8.get())
    a9 = float(e9.get())
    a10 = float(e10.get())
Button(master, text='Quit', command=master.quit).grid(
    row=30, column=0, sticky=E, pady=4)
Button(master, text='Show', command=lambda: make_ind_prediction(K)).grid(
    row=30, column=1, sticky=W, pady=4)
mainloop()
```

CHAPTER EIGHT

EVALUATION

We have applied various supervised algorithm techniques for the dataset; we have tabulated the value of accuracy, precision, recall, and confusion matrix for every algorithm respectively shown below:

Table (1.2) Tabulation for accuracy, precision, recall for various algorithms

The graphical representation shown below to have a better understanding of the accuracy, precision and recall we have achieved using various algorithms.

#	Algorithms	Accuracy	Precision	Recall	Confusion Metrix
-	Null	78	-	-	-
1	Logistic Regression	81.45	66.92	35.95	[8927 419] [1806 848]
2	KNN	78.86	53.47	34.17	[8557 789] [1747 907]
3	Naïve Byes	76.68	47.65	52.23	[7736 1610 [1188 1466
4	Classification Tree	78.46	52.09	32.81	[8545 801] [1783 871]
5	SVM	81.66	63.99	39.11	[8762 584 [1616 1038
6	Feed Forward NN	75.65	33.91	40.74	[8927 419] [1806 848]
7	Voting Classifier	83.95	67.49	32.83	[8842 504] [1822 832]

A. Precision-Recall Curve comparison

The below graphical representation PRC comparison of various algorithms. By comparing algorithms, a Voting classifier has good accuracy but when we draw PRC, it shows that Logistic regression has good Precision-Recall value at threshold 0.5. So, while changing threshold values, it improves the Precision and Recall values.

Logistic Regression Classifier to check threshold value: To check threshold value and Precision, recall values at different threshold, we draw Logistic Regression classifier diagram. Here, we shown good precision and recall value at threshold 0.2. So, updated a model with threshold value 0.2 and the improvement was approx. 44% in precision and recall value of a model. As, it decreases False Negative value which means defaulters are

predicted as non-defaulter. False Negative value is changed approximately 1800 to 1000 and the confusion matrix was

[7487 1859]

[1074 1580].

B. Graphical User Interface

We created Graphical User Interface using python and tkinter, we trained a model and set threshold value at 0.2 in logistic regression. When user will submit below mentioned parameters value, model will predict whether a user will be defaulter or non-defaulter next month in payment.

The general steps are mentioned below:

1. Choose the best model and parameters
2. Save to .json file
3. Load a file from disk to predict data
4. Call a function on button submit and load data to a model
5. Check probability and result on GUI

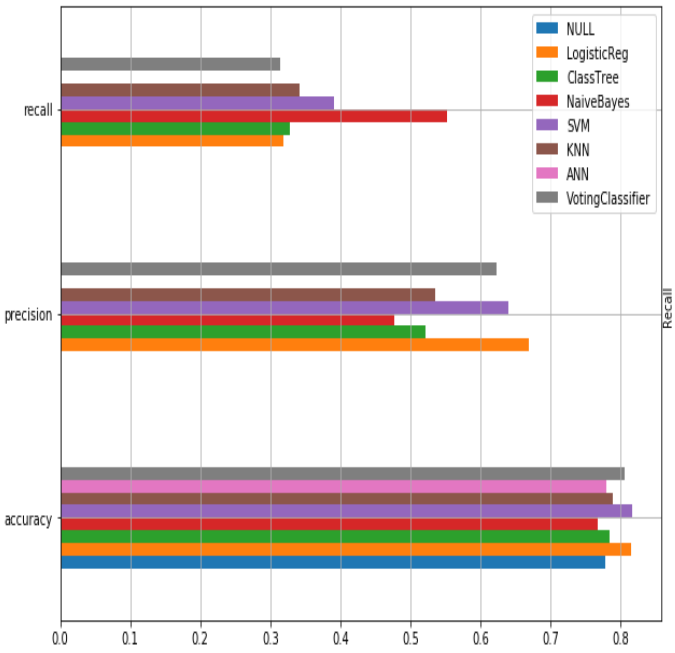


Fig (1.5) Accuracy, precision and recall for various algorithms

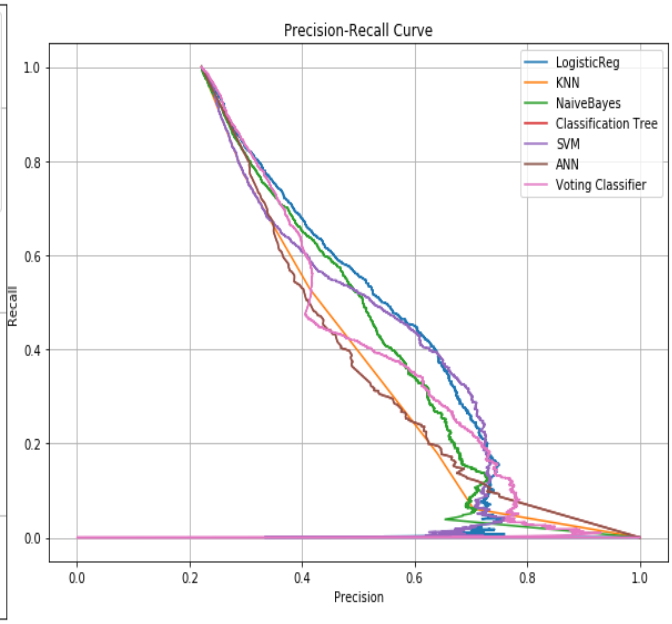


Fig (1.6) PRC comparison

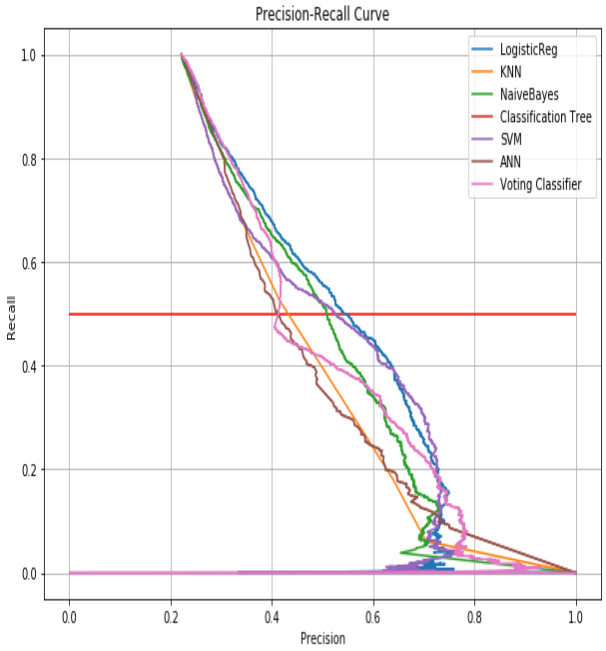


Fig (1.7) LR classifier

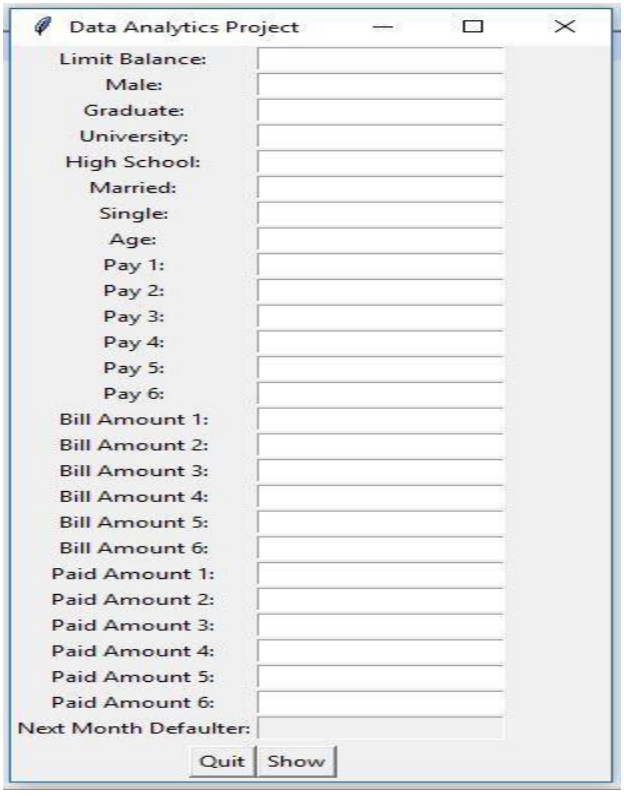


Fig (1.8) GUI for checking defaulters

CHAPTER NINE
CONCLUSION AND FUTURE WORK

This would inform the issuer's decisions on who to give a credit card to and what credit limit to provide. We investigated the data, checking for data unbalancing, visualizing the features and understanding the relationship between different features. We used both train-validation split and cross-validation to evaluate the model effectiveness to predict the target value, i.e. detecting if a credit card client will default next month. We then investigated five predictive models: We started with Logistic Regression, Naïve Bayes, SVM, KNN, Classification Tree and Feed-forward NN and Voting classifier accuracy is almost same. We choose based model Logistic regression based on minimum value of False Negative from confusion matrix.

REFERENCES

[1]. Li, Xiao-Lin, and Yu Zhong. An overview of personal credit scoring: techniques and future work. Journal: International Journal of Intelligence Science ISSN 2163-0283. 2012.

[2]. Yeh, I-C. and C-H. Lien, 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systerm with Applications,36: 2473-2480.

[3]. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

[4]. Taiwo Oladipupo Ayodele. (2010) "Types of Machine Learning Algorithms", New Advances In Machine Learning, Yagang Zhang (Ed.), Intech

[5]. NH Niloy, MAI Navid. Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients. American Journal of Data Mining and Knowledge Discovery. Vol. 3, No. 1, 2018, pp. 1-12. doi: 10.11648/j.ajdmkd.20180301.11

[6]. Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36, 3302–3308.

[7]. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves.ACM New York, NY, USA 2006. ISBN:1-59593-383-2.

[8]. Christopher M. Fraser(2000), "Neural Networks: A Review from a Statistical Perspective", Hayward Statistics.

[9]. Shie Mannor, Dori Peleg and Reuven Rubinstein. ICML '05 Proceedings of the 22nd international conference on Machine learning. ACM New York, NY, USA 2005. ISBN: 1-59593-180-5

[10]. Arlot, Sylvain, and Alain Celisse. A survey of cross-validation procedures for model selection. eprint arXiv:0907.4728. DOI:10.1214/09-SS054.