

LD7083

**Computing and Digital Technologies Project**

Supervisor Name:

Parvez jugon

**Dissertation Report**

**on**

**Identification of credit card defaults using machine learning**

**Student Name/ID:**

Makkena Adithya Sai Mahesh(19045379)

**Programme:**

MSc Computing and Technology with Advance Practice

Table of Contents

[Abstract 4](#_Toc92755967)

[1. Introduction 5](#_Toc92755968)

[1.1.Problem Statement 5](#_Toc92755969)

[1.2.Research question 6](#_Toc92755970)

[1.3.Significance of research 6](#_Toc92755971)

[1.4.Aims and objectives 7](#_Toc92755972)

[1.5.Constraints and limitations 7](#_Toc92755973)

[1.1. Structure of Report 7](#_Toc92755974)

[2. Literature Review 9](#_Toc92755975)

[2.1.Factors that influence credit card spending 10](#_Toc92755976)

[2.1.1. Credit card attributes 10](#_Toc92755977)

[2.1.2. Demographic factors 10](#_Toc92755978)

[2.1.3. Bank policies 11](#_Toc92755979)

[2.1.4. Attitude towards money 11](#_Toc92755980)

[2.2.Distinguish between good and bad customers 11](#_Toc92755981)

[2.3.Factors that lead to credit card defaults 12](#_Toc92755982)

[2.3.1. Total credit card debt to income ratio 12](#_Toc92755983)

[2.3.2. Size of the credit limit 12](#_Toc92755984)

[2.3.3. Inadequate awareness 12](#_Toc92755985)

[2.3.4. Proportional payment of expenses with credit cards 12](#_Toc92755986)

[2.4.Risk Default predictions 13](#_Toc92755987)

[2.5.Machine Learning (ML) to predict credit card default 13](#_Toc92755988)

[2.5.1. The use of the deep learning models 14](#_Toc92755989)

[2.6.Literature Gaps 14](#_Toc92755990)

[3. Research Methodology 16](#_Toc92755991)

[4. Data Analysis & Findings 23](#_Toc92755992)

[5. Discussion 36](#_Toc92755993)

[6. Conclusion, Future Study, and Self-Reflection 40](#_Toc92755994)

[References 43](#_Toc92755995)

**List of Figures**

Figure 1: Average Credit Card debts in top 10 countries 10

Figure 2: Accuracy Scores of Classifiers 37

Figure 3: Error/Loss Rate in of Each Classifier 37

Figure 4: Cohen Kappa Scores 37

Figure 5: Matthews Correlation Coefficient Scores 37

**List of Tables**

Table 1: Precision, Recall, F1-Score of Each Classifier 35

Table 2: Credit Card Defaults Prediction Results:

Accuracy, Loss, Cohen Kappa, and Matthews Correlation Coefficient 36

# Abstract

The use of credit cards is the fastest growing method of transaction because of its convenience, security and reward points. But the mode of credit card transactions also has raised debts due to credit card default payments. The defaults in credit cards happen in various forms that include application frauds, bankruptcy frauds, theft frauds, behavioural frauds etc. In recent years it is also witnessed that there is an exponential increase in credit cards frauds and thus it is required for the banks and financial institutions to predict the defaulters. The credit card portfolio for banks aims towards maximising profits as it helps to provide higher and more volatile returns for the banks thus requiring controlling the credit card default risks. The risk due to the credit card defaulters can be controlled by the use of machine learning and the development of a machine learning model to predict the defaulters. To implement the machine learning python program was used. The data was retrieved from the Kaggle datasets that contained both the dependent and the independent variables. The independent variable consisted of various factors that lead to credit card defaults that include demographic factors like age, sex, income and also credit card limits and payment limits. The process for analysing the data in Python included considering the various python libraries, understanding and cleaning the data like the use of customer ID without was not required. The process also included the use of StandardScaler as a pre-processing method, the use of various algorithms, Cohen’s kappa score and Mathew coefficient. The different types of algorithms that were used included Support Vector Machine (SVM), Logistic Regression and the use Decision Tree. Comparing these algorithms it is witnessed that the use of SVM generates the highest performance in terms of accuracy, recall, precision and f1-score. Thus through the use of the SVM algorithm, it helps to provide the best predictive model for credit card default payments.

# Introduction

Since the time the credit card was introduced in the 1960s, it has gradually replaced checks and cash transactions as a default payment mode of transaction. It is now the fastest-growing transaction method across the globe. The use of credit cards is now an opted mode of payment because of its convenience, security and reward point. The use of credit also does provide margins to many who may not be having sufficient precautionary savings (Lee, et al., 2019). But from the economic perspective, the change mode of payment do have an impact on increasing the debt which has been steadily rising over the past two decades. It is research out that although the national economy is healthy with an increase in per capita income, there is also an increase in the credit card payment problems accounting to payment defaults.

Payment defaults are one of the major ethical issues in the credit card industry. The extensive use of credit cards is the major issue of the credit card industry. Credit card defaults happen in various ways that depend on the type of fraud concerned. These may include bankruptcy frauds, theft frauds or counterfeit frauds. It also includes application frauds and behavioural frauds. Each of the types of frauds can be subcategorised having its speciality and definition (Delamaire, et al., 2009). Credit card frauds have witnessed an exponential increase due to the constant increase in the usage of the card but also due to ease to perpetuate the frauds. The aim is towards identifying the different types of credit card defaults and also to review the alternative techniques that are used towards detecting and predicting such defaults by the development of the model to predict credit card defaulters. Prediction of the credit card defaulters can be generated by the development of the model. The credit card default model is to be generated by the use of machine learning algorithms (Han, et al., 2020). There are different types of machine learning algorithms linear regression, logistic regression, Naive Bayes, kNN, Random forest etc.

## Problem Statement

Like any other industry the credit card industry aims towards maximising profits by measuring and controlling the risks and avoiding payment default as much as possible. The credit card portfolio of the banks provides higher and more volatile returns for the banks as compared with the other conventional product mixes (Odhiambo & Memba, 2012). The performance of the credit card portfolio also does affect the performance of the commercial banks and the financial institutions in terms of increasing customer satisfaction and also the revenue of banks.

The increase in usage of the credit card which is the alternative form of making payments and obtaining cash is one of the main portfolios in banking and also for financial institutions. But credit risk is one of the major risks for banks. The increase in competition to lend and the increase of retail bank expansion also have affected the risk of the banks as they do relax their credit approval and procedures to appraise a customer have increased the bank credit risks. The problem relates to inconsistency in the test standards they follow and also due to customer defaults. The banks are thus are required to understand the factors for defaults and can segregate between good customers and bad customers to make an effective decision for the credit card portfolio.

## Research question

The research questions include:

* What are the factors that influence consumers to spend more on credit cards?
* What are the factors that raise credit card defaults?
* How can banks and financial institutions be able to predict credit card defaults?
* How can the use of Machine learning algorithms be used to develop a model to ensure better predictions to reduce credit card portfolio risk?

## Significance of research

In the era of increased access to credit, it is important towards understanding the consequences to take unsecured consumer debts. Credit is a limit or a special form of money that is unique, fully convertible or exchangeable with other forms of currency. In short, credit is both a resource and also a liability that would require future payments to be made with interest (Hodson, et al., 2016). Thus it is risky and uncertain. These aspects are a part of the credit card business to carry out with the business of unsecured debts that leads to defaults.

Defaults happen when the borrower is not able to make convenient payments, or the consumer misses the payments, dodges or quits to make payments of the credit purchases. In the case of credit cards, there are no assets that would help the banks and financial institutions to recover the debts (Ala’raj, et al., 2021). Credit card organisations use to give a few months grace period of non-payment of dues before they declare as the default account. Thus the banks and financial institutions are required to develop a system for the organisation to predict the most likely credit card defaults.

The system is towards the development of a credit card risk model that would help to understand the probability of the customers likely to pay back. The model is to be developed by the use of machine learning algorithms. Manir, et al. states that the use of machine learning algorithms helps to analyse and investigate the transaction of the credit card holders, the history of credit and other types of data (Manir, et al., 2019). The use of machine learning algorithms and generating a model improves the credit card default detection performance over time.

## Aims and objectives

The research aims to develop a model to Predict the likelihood of credit card defaulters

for customers of the bank using the machine learning algorithms.

The objective of the research is

* To understand the factors that influence credit card spending among consumers
* To understand the factors that can lead to credit card defaults
* To develop a system for the purpose to predict credit card defaults that would help to reduce the risks.
* To develop a model with the use of machine learning algorithms that would help the bank to make effective decisions to reduce credit card defaults
* To complete the research and the development of the model as per schedule and timeframe and allotted budget.

## Constraints and limitations

The constraints and limitations of the research are concerning the formation of the research aims and objectives may be too broad as it can be narrowed down to the focus of the study of a particular bank credit card portfolio. The limitation of the study is also is related to the collection of the data as due to the Covid-19 situation no primary data gets collection. The research is mostly based on secondary data. The discussions that are covered in the literature review are less cited as there is a lack of previous studies in that particular area in the field of credit card defaults. Research constraints are also related to the scope of the research. The lack of experience that I have in conducting research and producing academic papers.

## Structure of Report

The structure of the document consists of 6 chapters for the machine learning data science project for credit card default. The chapters of the report are elaborated on below.

Chapter 1 consists of the Introduction of the report that helps to understand the research problem and specifies the aims and objectives of the research. The chapter also provides the research question, understands the significance of the research. The chapter also understands the limitations and constraints of the research topic.

Chapter 2 includes the literature review. The chapter helps to critically analyse the knowledge that is generated by the previous researchers. The topic in the chapter helps to understand the various facets of credit card default and also the machine learning aspects. The gap within the literature review is also specified.

Chapter 3 relates to the methodology that was used to carry out the research. The methodology considered understanding the machine learning process, understanding the attributes of the data that was used. The chapter also helped me understand the machine learning techniques and algorithms that were used.

Chapter 4 considered data analysis and findings. The chapter helped to understand the different steps for analyzing the data and its results and findings. The chapter specifies the codes that were used for analysis and the algorithms used.

Chapter 5 includes the discussion of the findings of the previous changes. The discussion included a performance of the different machine learning algorithms that were used and its comparison. The chapter is the best model for predicting credit card default.

Chapter 6 is the conclusion of the report helping to take note of the pain points of the report. The chapter also specifies future studies and learnings that are achieved from the dissertation project.

# Literature Review

The financial institution that includes the banks issue credit lending products that include credit cards, personal loans and mortgages. Such activities include the centre of the dealings and with proper lending yields the banks with huge gains. It requires these banks to get new customers and also ensure profits. The wide range of databases that the bank has generated over the years is used for analysing the performance of banks and making progressive business decisions (Ala’raj, et al., 2021). All customers of the banks won’t act the same way when it comes to financial performance and it is required for the banks to implement distinguishable treatment who qualifies for profitable requirements that are based on repayment and purchasing behaviour.

Consumer credit is regarded as a significant part of the banking system. The use of credit cards also do continue to be dominant and increasing popular modes of payment (Jantsch, et al., 2021). The card is issued to the users for the goods and services that are based on the holders promise towards paying them. The issue of the card does revolve around the account and the line of credit to the consumers that relate to the user borrowing money paying to the merchant or as a cash advance to the user. The downside of the use of credit cards is the tendency to default on the payment. The credit card system entails the greatest risk among all types of bank loans that increases the cost of operations that generates default risks.

In the US it is the second most popular non-cash instrument. Teoh, et al states that the use of credit cards becomes a convenient way to expand purchasing power (Teoh, et al., 2013). It allows the consumers the flexibility to defer payments to a future date and thus it allows consumers to smooth spending over temporary liquidity shortfalls. It helps to ensure convenient transactions and is also easy to carry around but the disadvantage lies in overspending and it also comes with interest. The percentage of owing the general-purpose credit card includes MasterCard, Visa, Discover which is 70.2% in American houses that are based on the latest US Censor Bureau. In 2019 the US Credit card debt reached $971 billion which represented a 5% increase over 2018. It reflects an upward trend (Ma, 2020). The rise in the credit default rate in the UK adds concerns that increased to 22.9% in the first quarter of 2019 from 12.7% in the last quarter of 2018 (Inman, 2019).

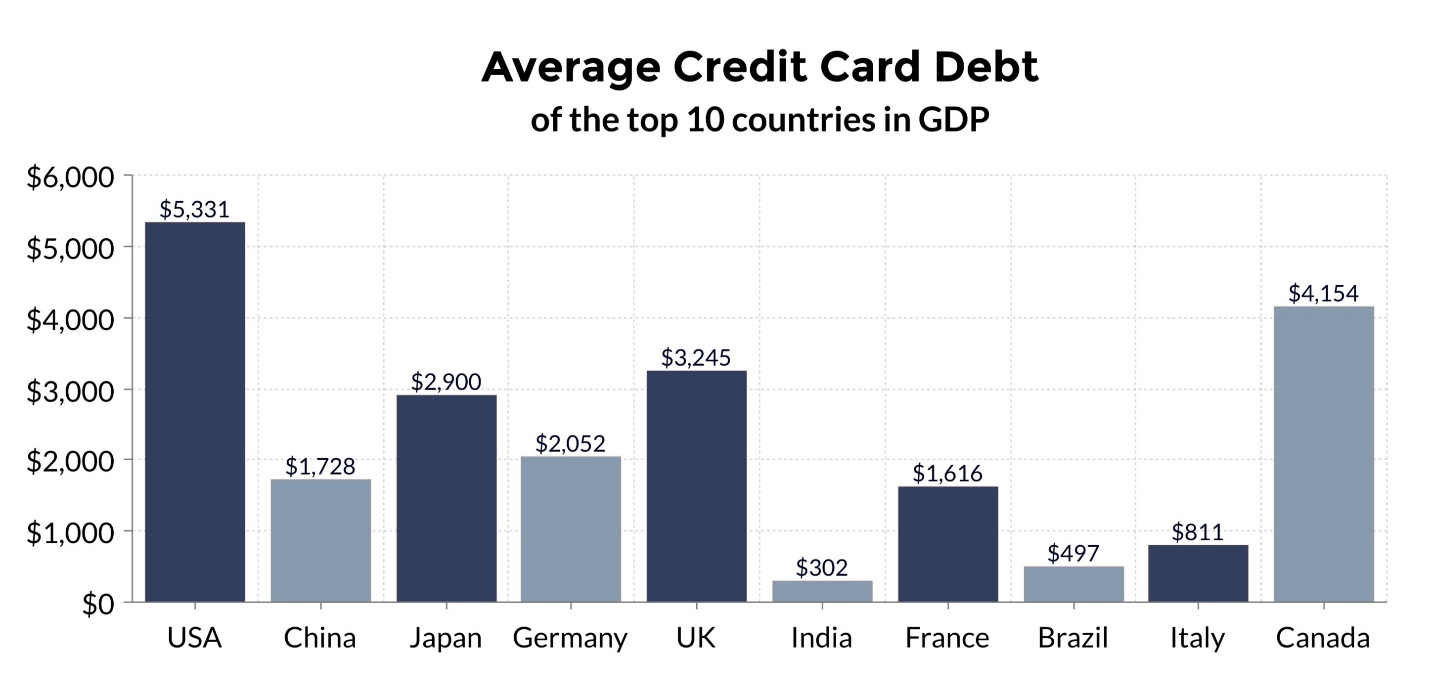


Figure 1: Average Credit Card debts in top 10 countries (Shift, 2021)

## Factors that influence credit card spending

Various factors influence credit card spending. Evidence does suggest that credit cards do take advantage of cognitive bias and another psychological mechanism. In many cases, consumers do overestimate their future ability to repay and are also surprised by the interest charges when due happens. Studies reveal that shopping with the use of credit cards leads to consumers being willing to spend more on items and they check out having bigger baskets (Banker, et al., 2021). The focus is on more product benefits rather than remembering the cost.

### Credit card attributes

The study and the impact of the credit card attribute that results to the debts are the results of lifestyle influence. The attributes also relate to aspects of self-esteem, time consciousness, pressure from peer groups, the influence of advertisement. It relates to convenience, use patterns and the status of the person (Khare, et al., 2012).

### Demographic factors

The demographic factors include aspects like age, gender. It also relates to different other aspects like self-fulfilment, a sense of belonging and the aspect of security. Teoh, et al states that the age of the holder acts as a very important factor (Teoh, et al., 2013). Old customers are more likely to hold one or two credit cards as compared with college students or young credit cardholders who may have fewer credit cards but accounts for increased credit card debts. It is related to the perception of the credit card holders as they believed that they do have a long future to settle the debt and also assume that the financial problems are temporary. There is also a significant relationship between the income level and the credit card holder as it affects the consumption behaviour of the use of credit cards.

### Bank policies

Due to the high competition that exists in the credit market, the banks do thrive to establish their niche market through constant improvement of product and innovation. Some various banks and non-banks offer different incentives that include cash rebates, no annual fees, instalment payment plans, and discounts for identified purchases. The desire to spend more on credit cards are enhanced, acceptance of major credit cards etc. These factors do influence the credit card spending of the customers.

### Attitude towards money

The cognitive potentiality of the attitude towards money represents a person's achievement in society, relates to a symbol of success. The aspect of money also helps to gain respect in society as the symbol of identity. It also reflects a sense of freedom or to provide power to the individual (Yayar & Karaca, 2010). The defined attitude towards money does impact the area of the life of the person that includes saving, spending and the attitude that includes power prestige. It also reflects the willingness to pay the credit card expenditure and the awareness about interest charges and total debt owed.

## Distinguish between good and bad customers

It is required for the banks to distinguish between good and bad customers. To distinguish customers, the customers are required to be credit scored and behavioural scored. It is through credit scoring that helps and analyse the likelihood of the applicant to repayment falter or likely to pay. The aspect of credit score refers to credit which means to buy an item and pay afterwards and scoring is a method that is used for credit cards (Ala’raj, et al., 2021). Credit scoring is of two types which include application credit scoring where the score is applied to provide decisions on new credit applications and behavioural scoring that helps the banks to guide decisions related to lending to the management of credit card limits. This helps to manage the collection and recovery of debts. It also helps to retain future profitable customers and to predict accounts likely to close and ensure early settlement. Moreover based on the credit score the banks also offers new financial products and rate of interest. It helps to manage dormant accounts, optimize the telemarketing operations and to predict fraudulent activities. It helps to ensure several risk payments and future risk payments.

## Factors that lead to credit card defaults

### Total credit card debt to income ratio

The issue of the credit card is required to take into consideration the total debt to income ratio which is regarded as an important aspect to understand the financial health of the customers. It is required to calculate such aspect that depends on the current debt situations and how comfortable the customers are in regards to making the payment of credit bill. This will help to understand the risk association while taking to another payment.

### Size of the credit limit

The banks are required to take note of the customers who can be trusted and towards understanding the credit limit. The decision to allocate the credit limit is a major source of dilemma that results in several credit card defaults. The increase in the credit limit does generate a significant size in the consistency of debt. The factor is also related to the interest-elasticity of credit card limits. The banks to strategize against Ponzi schemes tends to provide lower credit limits to borrowers having high risks even though they issue a large number of cards.

### Inadequate awareness

The intuition of the people to take note of the payoff and towards understanding the consequences for late payment and financial consequences. The customers are required to note the payoff times and towards understand the consequences of late payment and financial burdens. The factor that would influence the understanding of the customer in debt is how the information gets reported in the monthly statement (Soll, et al., 2013). As per the policy of the government, it is required to include a table of how long it will take towards paying the balance. Customers are also required to be informed of the minimum amount that is required to be paid each month and the duration of the payment. In most cases, people do not have the correct formula to solve the problem.

### Proportional payment of expenses with credit cards

Credit card defaults arise with the proportional to consumption that would limit income and the other expenses that are incurred by the consumers. If the expenses that are to be incurred by the others are high, it becomes difficult for the consumers to repay the debts.

## Risk Default predictions

The banks are also required to predict when the customers fail to pay the credit card bills. They are required to calculate the profitability over the lifetime of the customers using profit scoring. They are also required to ensure the average of default levels over time and to ensure beneficial to ensure debt provisions (Ala’raj, et al., 2021). It is also required for them to assist the terms and also to adapt to change in economic conditions. It is also required for the bank towards estimating the credibility of the borrower and to provide a safe probability when the customers are more likely to miss a payment. The use of the models helps banks towards taking actions quickly that helps to reduce the risks from credit card defaults.

The use of the default analysis system allows the bank to monitor and eliminate the ongoing transactions of the potential credit card holders as risks before occurrence and defaults. There is various risk prediction system that helps to detect credit card defaults that are based on statistical techniques. Chou & Lo predicts that there are studies that are initiated the method of the use of Artificial intelligence system performs better as compared with the use of statistical methods for assessment of credit risks (Chou & Lo, 2018). Machine learning and data mining techniques are used that helps to assist data managers towards improving the problem-solving technique. The use of these techniques helps to discover meaningful patterns and to understand trends from large amounts of data

## Machine Learning (ML) to predict credit card default

Machine Learning and deep learning that include image recognition, natural language processing, detection of anomaly and robotics are providing a mainstream solution that would help to improve the accuracy of assessing credit card risk. Muni & Jung states that ML is a kind of artificial intelligence that programs to set rules in computers but rather allows the computers towards generating the rules that would help to perform a specific task by using self-learning given data (Muni & Jung, 2021). It is ML that helps the system perceive a certain situation as per experience. ML is a process towards discovering the rules or patterns through learning from different types and sources of datasets.

The development of a robust ML model helps banks to decide about credit applications and at the same time ensure that clients be aware of the behaviour that damages the credit score. The main motivation to come up with risk prediction is to utilise financial data. These data include client transactions, exchange records, business transactions records etc. that help the banks to decrease loss and vulnerabilities. Several prediction models are used that includes logistic regression, nearest neighbour, discriminate analysis (Butarua, et al., 2016). Yaseen, et al stated that to assess the credit card default risk six different machine learning models are used (Yaseen, et al., 2020). These models include Naïve Bayes, Bayesian network, random forest, J48, multilayer perceptions and logistic regression. It evaluates precision, accuracy and recall. The models that are generated are compared to forecast the default across banks that helps to ensure better risk management. The model also provides the ability to compare drives across different banks.

Other different fraud detection techniques include the K-Nearest Neighbour algorithm that helps to determine anomalies in the target instance. The hidden Markov model is also used to detect fraud activity at the time of transaction. The use of a Neural Network after learning from the previous behaviour can detect real-time credit card frauds (Pumsirirat & Yan, 2018). Decision trees are used to handle non-linear credit card transactions. The use of the outliner detection method helps to detect credit card fraud having less memory and computing requirements. The use of deep learning helps to analyse and also ensure learning of the massive amount of data that is unsupervised.

### The use of the deep learning models

The use of the conventional neural network that is built with the use of multiple hidden layers do solve various image classification and towards recognising the problems in scenarios that are complex (Chou & Lo, 2018). Each of the layers that are hidden learns from the previous layer abstraction. There are deep learning algorithms and models that are used that includes deep neural network (DNN), the use of the convolutional neuron network (CNN), the use of autoencoder (AE) etc. The problem for the developer lies to select the algorithm towards solving the problem and understanding the aspect of each of these algorithms does. The developer is required to know the real problem that helps to ensure effective selection of the algorithm to use in deep learning unsupervised learning (Pumsirirat & Yan, 2018). The use of unsupervised learning does automatically extract meaningful features of data that are not labelled and ensure training regularisation.

## Literature Gaps

The literature lacks default analysis taking into account the credit card portfolio. Gaps lie in producing the estimates for the customer can afford depending on the number of credit cards the customer holds. The limits of the credit card which is determined by the credit score of the individual thus lack the aspects of deciding the credit limit of the customers based on different factors like income, expenses, living standards etc. are not considered. The literature also witnesses gaps towards understanding the implications of the collective factors of the use of credit cards. There is hardly any customer segmentation clusters being established with the most valuable customers who would be generating more business. Gaps also exist of the factors that influence credit card users like the interpersonal factors are less studied. The aspect of social influence and individual desires is not covered in the literature that helps to reduce the perception of the use of credit card risk that would help to increase the acceptability of the use of credit cards in developing countries. In such regard, the element of trust and security of the credit cards are missing from the literature that would help to reduce the perceived risk.

# Research Methodology

**Introduction**

This chapter explains the methodology used to build a machine learning system for predicting the defaulters based on the aims and objectives. this chapter also explains about the data selection, machine learning algorithms adopted for predicting the defaulters.

**Research modelling**

In the current research we are using the crisp modelling. **CR**oss **I**ndustry **S**tandard **P**rocess of data mining is a process model with six phases that describes the life cycle of data science. It’s a set of guardrails to help us plan, organize and implement our data science (machine learning) projects

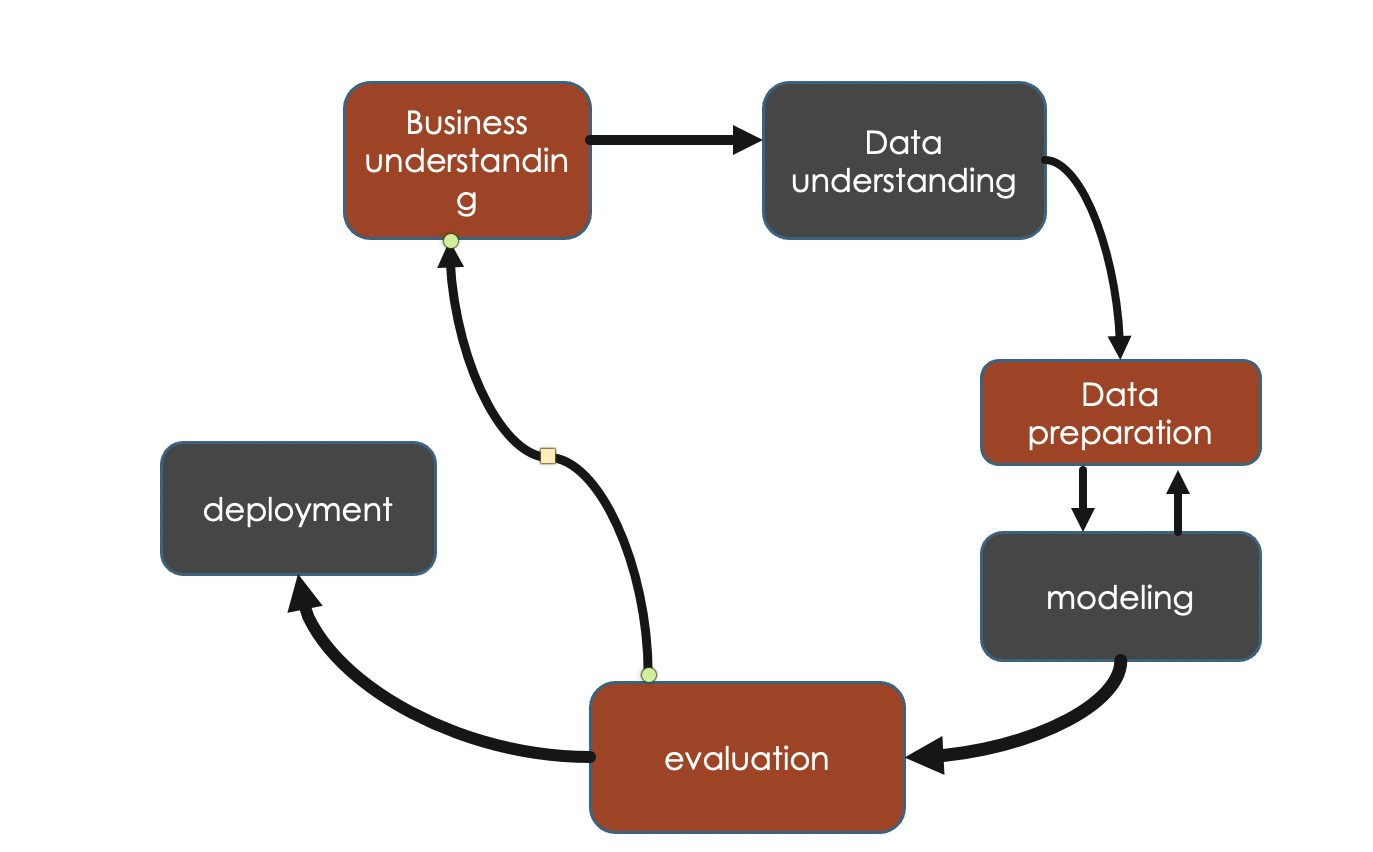


Figure 2 **Crisp modelling**

Business understanding:

The main objective for the research is designing a model to predict the credit card defaulters using machine learning algorithm using machine learning we could train the past data and predict whether the customer may defaulter or not using different algorithms.

Data understanding:

In this project we have a secondary data from the Kaggle, In the data set we have various attributes like age, sex, past billed amounts, and past paid history so that depending upon the various attributes in the data set.

Data preparation:

The data must be cleaned in order to perform ML algorithms i.e., eliminating the null from the data set and dividing the data into training and testing parts and applying standard scaling.

Modelling:

Applying the various machine learning algorithms to the trained attributes and predicting the customers .in the current research we are using logistic regression, decision tree and support vector classifier.

Evaluation:

Calculating the accuracy in each case for all the three algorithms and evaluating the ideal system for predicting the credit

**Research approach**

In this research we adopted quantitative approach by the selecting the past credit card customers details and using various machine learning algorithms to predict the new applicants and also qualitative approach in selecting the model by calculating the accuracy of the obtained results.

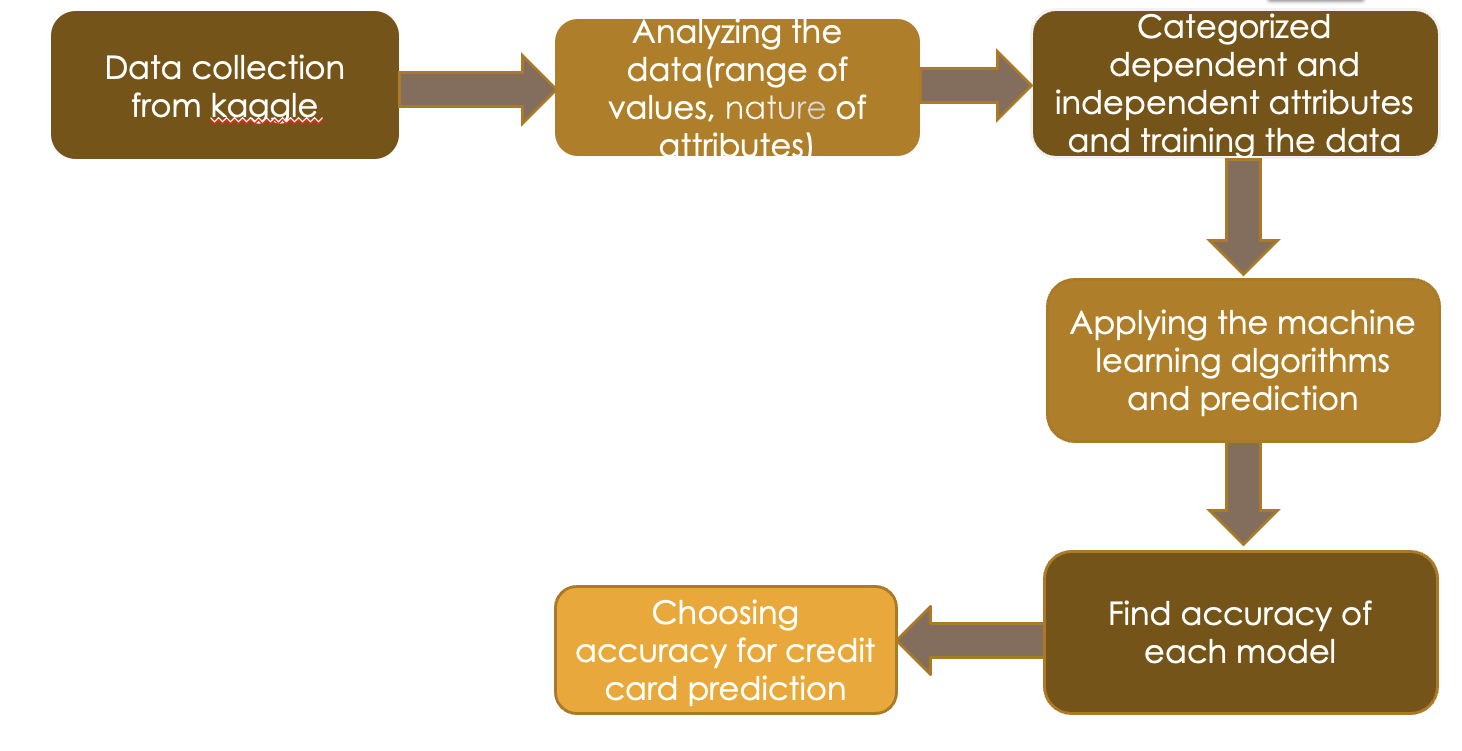


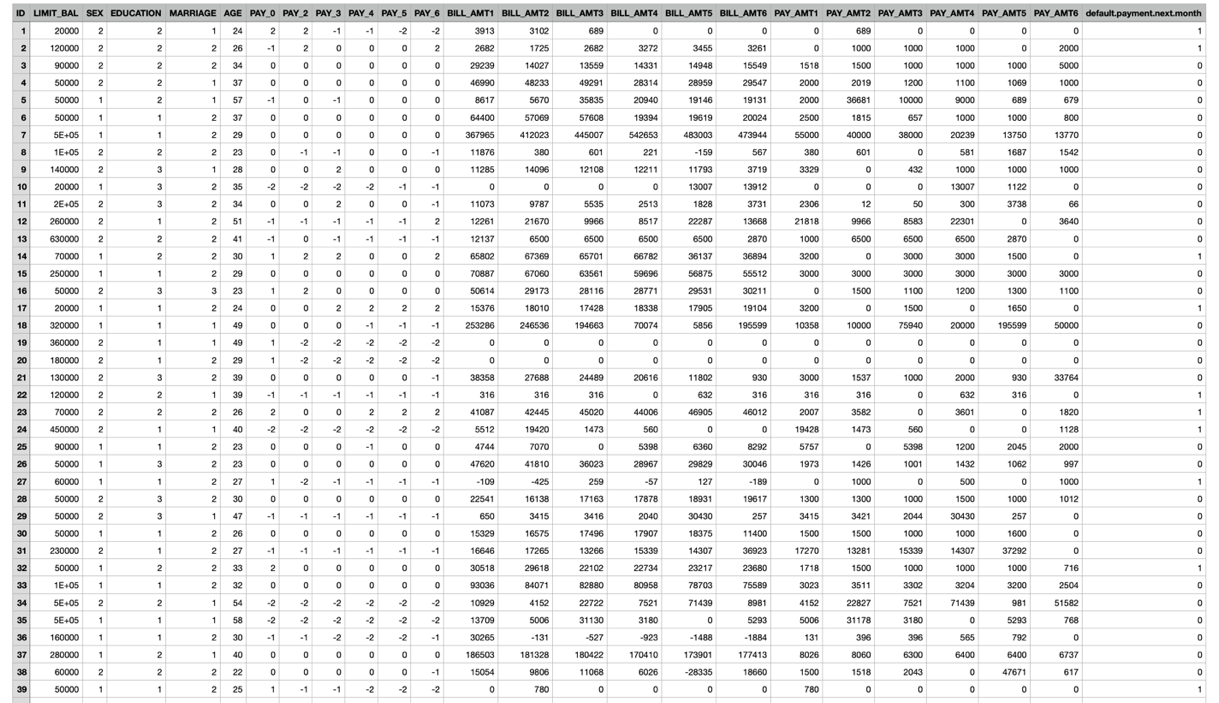
Figure 3: Flow chart of the research process

This project involves various stages:

1. Data selection and performing the data cleansing techniques and analysing the data based on the obtained data.
2. Implementing pre-processing algorithm and training the data based on dependent and independent variables to perform predictive models.
3. Performing the proposed machine learning algorithms and predicting the customers.
4. Finally, the next process is towards model evaluation as per the level of accuracy in the testing part of the data

**Data selection and explanation of datasets**

The dataset was selected from the Kaggle database. The dataset contained the “default payments of the credit card clients in Taiwan from 2005” that helped to conduct the predictive analysis. The data is included in the UCI\_Credit\_Card.csv file containing the data. The datasets do contain information about default payments, credit history, and demographic factors, history of the payment and the bill payment and bill statement of credit card clients from April 2005 to September 2005. The data consisted of 23 independent variables and a dependent variable of default payment. The dataset also contains the dependent variable the default payment next month which is represented by 0 and 1. The number of rows of the data includes 3000 rows. The data which we are using in the research is secondary data of UCI\_credit\_card .

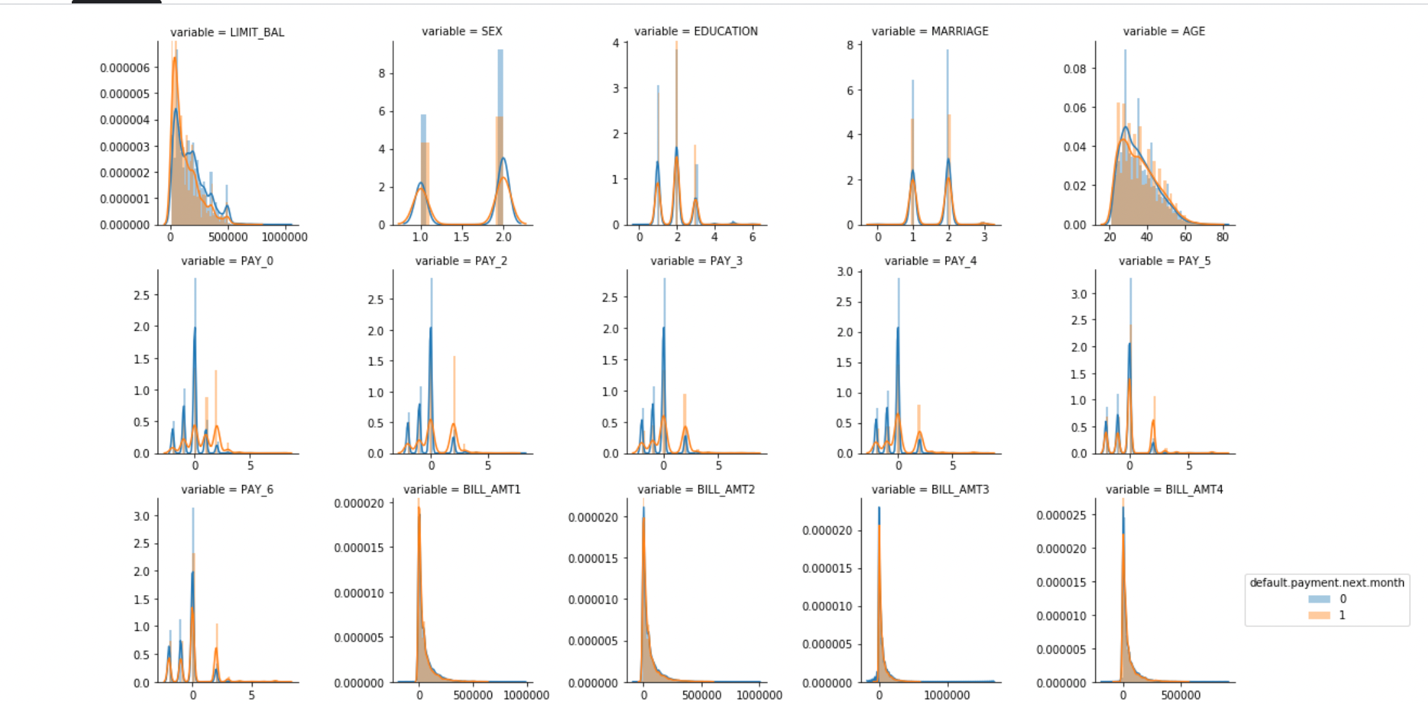


Data source: <https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

**Explanation of the attributes of data**

The datasets include 23 independent variables that include the limited balance of credit card, education qualification, sex, age, marriage, different pay levels and the bill amount that the customers needed to pay. These are all the input columns. The type of data type is quantitative data that includes discrete data which cannot be subdivided based on nature. These are in the form of a ratio in the form of pay or the salary of the credit cardholders. The variables used in the data set are

* Limit balance
* Sex
* Age
* Marriage
* Pay (0-5)
* Bill amount (0-5)
* Paid amount (0-5)
* Default payment next month



**Observations**

* Females are having the higher proportion of non-defaults(sex=2)
* More educated are having the higher proportion of non-defaults (education =1or 2)
* Singles are the having proportion of non-defaults(marriage=2)
* Age group of 30-40 are having higher proportion of non-defaulters

**Use of different algorithms**

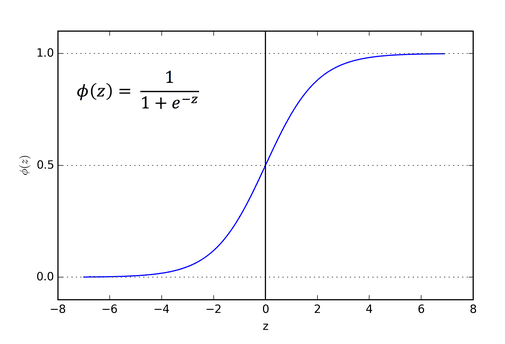
Analysing the credit card default data is required to take considering the use of different machine learning algorithms.

**Logistic Regression**

In ML supervised learning technique logistic regression is the most common approach that helps to predict and helps to categorise the dependent variable default payment-next month by the given set of independent variables.

In this research the output is either defaulter or non-defaulter .It is through logistic regression that would help to predict and categorise dependent variables with a yes and a no giving a probabilistic value that lies between 0 and 1.

0≤hθ(x)≤1

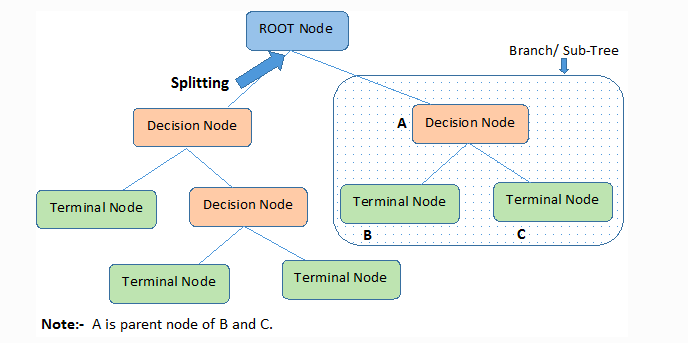


We used sigmoid function as our prediction value lies between 0 to 1 i.e. defaulters and non defaulters

**Decision tree classifier**

In supervised learning techniques, a decision tree is used for the classification of the datasets. Decision tree classification is also used for regression problems but in most cases, it helps to solve the classification model. It uses the CART algorithms. The tree-structure classifier includes the internal nodes that representations the features of the dataset that represents the decision routes and each of the nodes do represent the outcome. The use of decision tree classification does have two nodes that include the decision nodes and leaf nodes. The decision nodes are used for deciding with having multiple branches.

The decision is performed as per the dataset UCI\_Credit\_Card.csv as it gets represented through graphical representations to get all the possible solutions to credit card default payment.



**Support Vector Machines (SVM)**

The use of support vector machines is used for classification. The goal of the SVM algorithm is towards creating the best decision making that helps towards segregating the n-dimensional space of the data into different classes. This helps to put it in different data points that help to provide correct data categories. It creates a decision boundary that is referred to as a hyperplane. SVM is used for choosing the extreme points or vectors that help in creating the hyperplane. UCI\_Credit\_Card.csv uses the line at SVM that helps towards separating the data into two classified classes the credit card defaulters and credit card non-defaulters. SVM uses a single line that helps to separate the data and ensure data classification.

**Naïve Bayes**

Supervised learning is based on the Bayes theorem that is used to solve a classification problem. It is referred to as a probabilistic classifier that means it helps to predict the basis and the probability of the objects. The use of this algorithm helps to assume the occurrence of certain features that are independent of the occurrence of the other features that help to identify based on pay and bill amount. Each of the features contributes to the identity that would help depend on each other.

**Hypothesis Testing**

The ML project to generate a model for credit card default and classify the different criteria that lead to credit card defaults that include age, sex, pay, credit card limit etc. It is by undergoing the hypothesis testing that would help to ensure if the model is required to be refined. Thus it helps towards making the statistical decision by the use of experimental data. It is an assumption that we made about the parameters of the population that is supported by the sample data. The testing helps to ensure better testing of the exploratory data. It is testing based on the known sample measurement and with the use of various testing like Z-test, T-test, F-test, Chi-Square test. It is done towards conforming the observation about the population of the use of sample data within the desired error levels. It is hypothesis testing that helps to determine to conclude the hypothesis about the population is true or not. The steps that were initiated to perform the hypothesis testing was first to formulate the hypothesis, determine the level of significance, determine the type of test, calculation of the test statistic values and the p-value. The level of significance of testing is to be determined as per the level of significance. The significance level is set as 0.05 with having the chance to accept the null hypothesis. The computation of the P-value with less than 0.05 will help to reflect the behaviour and towards calculating the t-statistics towards deciding the acceptance and the rejection of null-hypothesis.

**Correlation analysis**

The variables in the dataset UCI\_Credit\_Card.csv are correlated that can be can relate to the other variables like one variable may cause or depend on the values if the other variable or slightly associated with the other variables or that two variable may depend on the third variable. It is correlation analysis that helped towards understanding the relationship between the two variables. The correlation that was determined understanding the relationship between various variables was in the positive correlation, negative correlation and neutral correlation. The correlation analysis was carried out by importing the pandas' library and then getting the datasets from the seaborn library that helped to load the dataset and then to get the spearman correlated coefficient with the method of spearman. By analysing the correlation in ML acted as an important tool that helped to build up the ML model. The use of the correlation analysis helped to take note of the variables that are strongly related to the machine learning model making it much simpler for interpretation of the data. The correlation that was conducted helped towards visualisation of the relationship between different variables.

**Cohen’s Kappa Score**

Cohen's Kappa score relates to the qualitative measurement of reliability for the two rates and the ratings are to be corrected of how the rates are required to be agreed (Warrens, 2015). It helps towards assessing the nominal agreement that is between the two rates with the value of 1 with having the perfect agreement between the rates and the value of 0 which means a random agreement. The aspect of the score has been interpreted between the value of 0 and 1. It helps to understand the aspect of reliability and validity of the data.

**Linear Regression Model**

The method of analysis for the credit card default to take into consideration the linear regression model is based on the ML supervised learning model. It is through the linear regression model that helps to understand the relationship between the input variable and that of forecasting. It is through Linear Regression that helps to relationship considering the cause and effect. It has helped to understand the types of relationships between the dependent and the independent variables. Thus it helped to understand the relationship between age and credit default payment next month, the relationship between education and credit default payment next month (Pedro Larrañaga, et al., 2018). The use of the ML model takes into consideration training data and the level of supervised data. By training through such a model helps to find the value of interception and also the value of the coefficient. It is through the value of the coefficient and value of regression that helps to achieve the best-fit regression line that helps to predict dependent value. It helps towards understanding the square root mean errors that help to achieve the best fit.

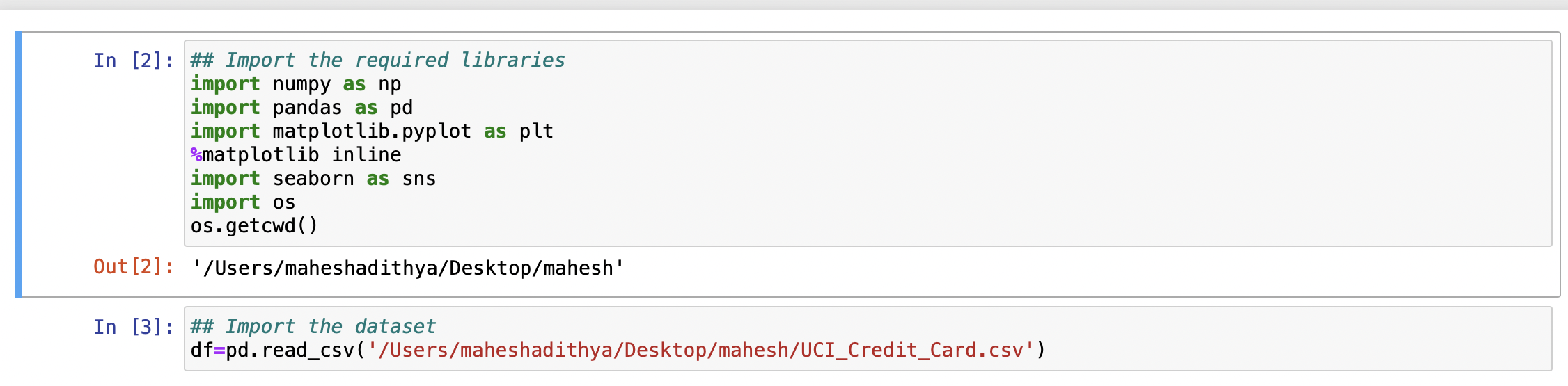
**Confusion matrix**

A confusion matrix is a table that is used towards defining the performance of classification algorithms. It helps to visualise and summarise the performance of classification algorithms. The confusion matrix is calculated in python by taking into consideration sklearn. matrics library. This takes into consideration the factors for credit card defaults (Visa, et al., 2011). It also helps towards understanding the effectiveness of the ML model and helps towards measuring classification. The use of a confusion matrix considers aspects like a true positive, true negative, false positive and false negative and with combination helps to predict actual value. It is through True Positive that helps to reflect and towards interpreting the positive that is true. It is True Negative that helps to interpret what is negative that to be true. Type 1 errors are predicted as false positive that helps to predict the positive that is true, The False True are the Type 2 Errors that helps to predict the negative that is false. It is the use of a confusion matrix that helps towards understanding the accuracy of the model. The model with having 99% of accuracy is regarded to be an excellent model and the best model to understand the credit card default project.

# Data Analysis & Findings

**4.1 Importing python libraries and data set**

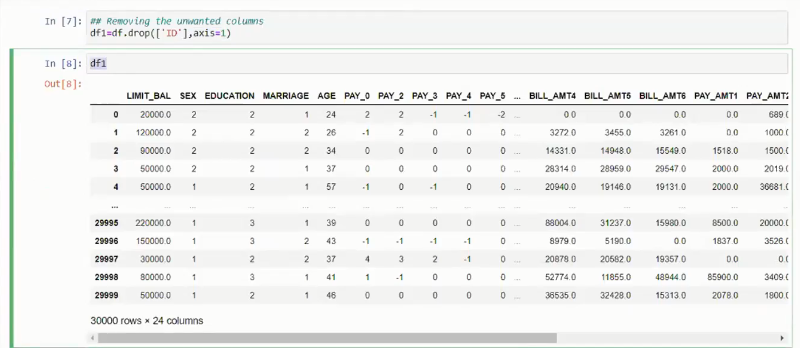
In the first step, some required python libraries are imported into the code and it is shown in the following figure. The required dataset needs to be imported into the data in the second step and for that, read\_csv function is used by placing the path of .csv file.



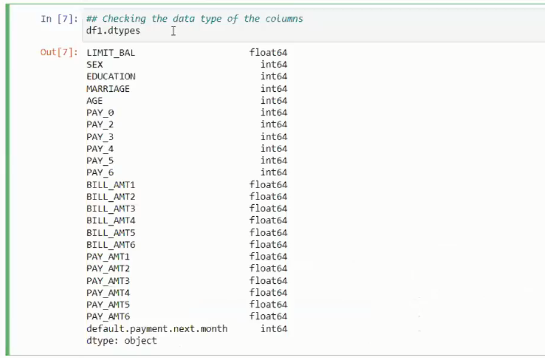
The data that is imported is stored inside the data frame defined and the imported data is tested by using the df.head().

**4.2 DATA CLEANING AND DATA ANONYMISTION**

The next activity is eliminating the null values from the dataset and to do that df.isnull().sum() command is executed and the results showed the zero null values for all the columns. The dependent variable from the dataset is the target column which is “default payment next month”. The rest of the columns except ID column are considered as the independent variables from the dataset. ID column only includes the list of numbers in order which does not make any impact on the default payments column. So, ID column is removed by executing the command of df1=df.drop([‘ID’]).axis=1). In that, axis =1 indicating matching to delete the entire column having the name of ‘ID’ and the results are given below and it showing that dataset is having 3000 rows and 24 columns.

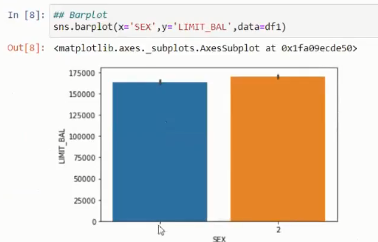


For implementing the machine learning algorithms, all the input data should be in terms of numerical only and to achieve that, it is required to verify the data types of all the columns from the data set. The command to verify the data types is df1.dtypes and it returns all the data types which are shown below. The results indicate that all columns are in terms of integers and float which are numerical values and so, all values can be used directly to implement machine learning without any issues.



* 1. **PLOTTING THE DATA**

The next activity is generating the plots and graphs from the dataset to analyze the data. Libraries such as matplotlib.pyplot and seaborn are used for that. The following figure shows the barplot created by considering the column of ‘SEX’ and 1 and 2 are indicating male and female values. Seborn is used to display the columns in different colors which is easy to visualize.



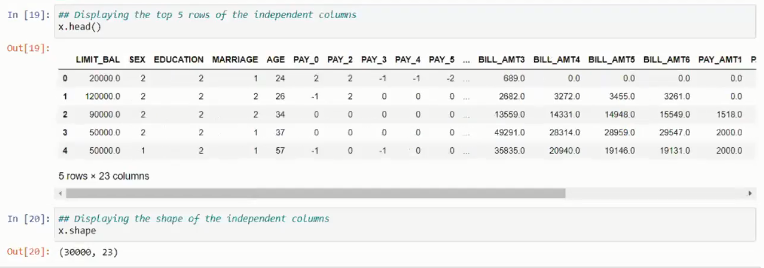
To count the values, sns.countplot() is used and then, warnings function is imported to eliminate the warning while generating the plots or executing different commands. Then, the distribution graph is generated to visualize the average or density of the columns from the dataset. Afterward, a box plot is created by considering the ‘MARRIAGE’ column and the results helped to analyze how the values are spread in a specific range of values. All these plots helped to have a good understanding of data so that algorithms can be implemented efficiently in further activities.

**4.4 SEPERATING DEPENDENT AND INDEPENDENT VARIABLES**

The next activity is mentioning to the machine about the independent and dependent variables. In the dataset, only the “default payment next method” column is a dependent variable and so it is removed and placed in a separate variable to simplify analysis, and commands used are given below. ‘x’ is storing all the columns after dropping the column “default payment next method” whereas ‘y’ is storing only that dependent column.



The following figure shows the results of x.head() and the shape of the x variable. Now, it has a total of 3000 rows and 23 columns.

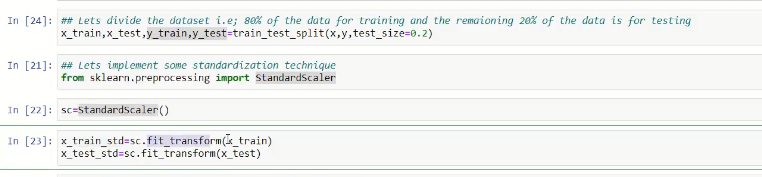


The following figure shows the results of y.head() and the shape of the y variable. Now, it has a total of 3000 rows and 1 column.



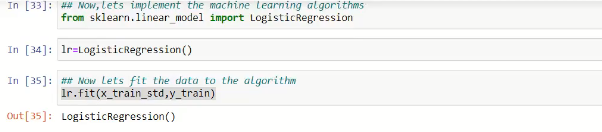
**4.5 DIVIDING TRAINING AND TESTING PARTS**

The next activity is dividing the dataset into training and testing parts. It is required to do for both the input and output columns that are identified above. The library used for that is train\_test\_split and it is from sklearn.model\_selection. In the project, 80% of the data is used for training and 20% of the data is used for testing. The variables defined for training and testing are x\_train, x\_test, y\_train, y\_test. train\_test\_split(x,y,test\_size=0.2) is used for this purpose. 0.2 indicates the ratio (80:20) adopted to split the data. From the x, 80% of the data is stored in x\_train and 20% of data is stored in x\_test. Similarly, y\_train stores 80% of data, and y\_test stores 20% from ‘y’.



**4.6 IMPLEMENTING MACHINE LEARNING ALGORITHMS**

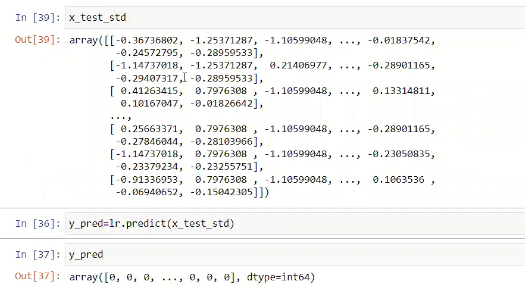
The next activity is implementing preprocessing techniques and it must be done only on the input columns. In the project, Standard Scalar is imported from sklearn. preprocessing and it is one of the standardization techniques. The Standard Scalar variable sc is used to fit the x\_train variable and is assigned to x\_train\_std and in the same way, x\_test is fitted to the variable called x\_test\_std as shown above. fit\_transform() function is used to perform those activities. x\_train\_std and x\_test\_std are now considered variables with preprocessing data. The major part of implementation is adopting machine learning algorithms and it starts with Logistic Regression. The class sklearn.linear\_model is used for importing the LogisticRegression algorithms into the code. A fit function is used to fit the variables of x\_train\_std and y\_train with the help of LogisticRegression variable lr.



The training part is implemented using the algorithms. The output columns and test dataset is used for analyzing the accuracy stores.

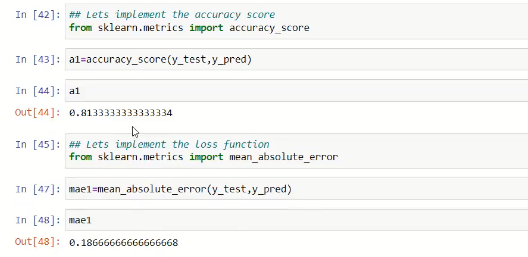
x\_train\_std and y\_train are used for implementing the machine learning algorithms.

Whereas x\_test\_std and y\_test are used for analyzing the accuracy scores. Usually, the accuracy store is identified by using the output columns and test dataset. Here, the test data set is y\_test and it is required to find the output columns from x\_test\_std and it is done by using the predict function and the command executed is y\_pred=lr.predict(x\_test\_std). The command indicates that input columns x\_test\_std are passed using predict, lr functions and are stored in y\_pred which are output values stored in the form of the array which are shown below. Now, these y\_pred and y\_test values are used to find out the accuracy.

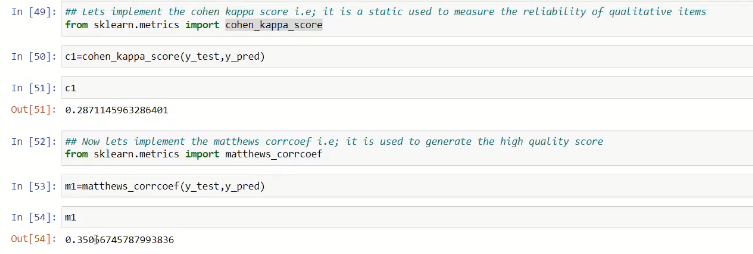


**4.7 TESTING THE ACCURACY**

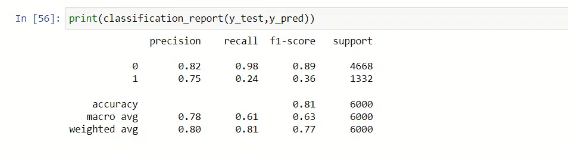
The next activity is importing the accuracy\_store package from sklearn.metrics and passing y\_test and y\_pred values into it as shown below. The results shows that logistic regression gives an accuracy value of 81.33% and it is passed into variable a1. Then, the loss function is implemented using mean\_absolute\_error and the error rate is 18.66%.



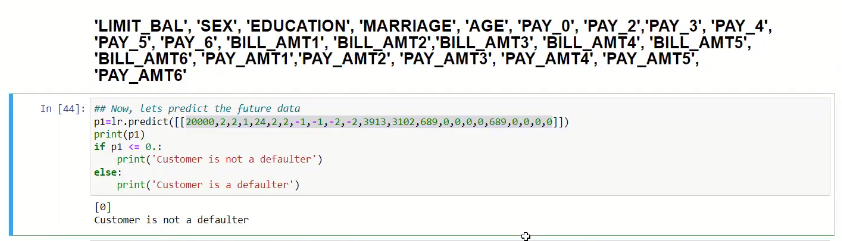
In the next activity, the reliability of qualitative items is measured by using the cohen\_kappa\_score function from sklearn.metrics and is shown. The function helps to give an idea about the quality of data used for the analysis and got the result of 28% which indicates a good score. Then, the Mathew corrcoef function is used to generate the high-quality score and the results show 35%. The commands used and the results obtained are shown below.



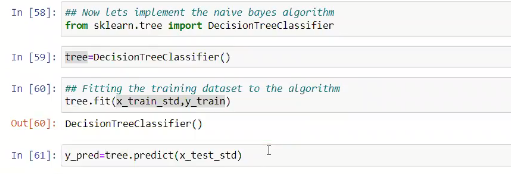
Implementation of classification report is the next activity in the process and classification report is imported for that purpose and passed the values of y\_test and y\_pred. The classification report includes some important details of precision, recall, and f1-scores for each 0 and 1 classification separately.



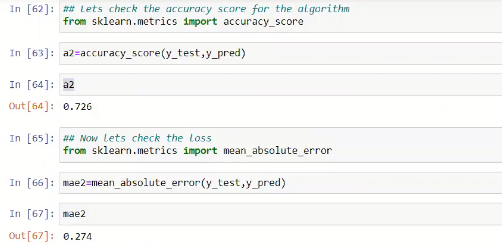
The process is followed by predicting the future data which is an important concept and it is done by passing the numerical data from the columns into the predict function with the help of logistic regression and storing the results in the p1 variable which is shown below. For the input data, the logistic regression predicted it as ‘0’ which indicates that the “Customer is not a defaulter”.



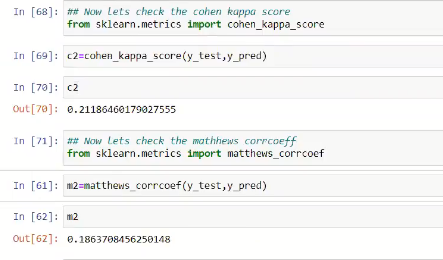
After predicting the defaulter using logistic regression, the decision tree algorithm is implemented. To achieve that, DecisionTreeClassifier is imported from sklearn.tree model and the same process used for logistic regression is followed. The first step is fitting the training data sets such as x\_train\_std and y\_train using a fit function with the help of the decision tree variable ‘tree’. Same as the previous one, y\_pred values are obtained by passing the x\_test\_pred into predict function and this time, the decision tree algorithm ‘tree’ is used.



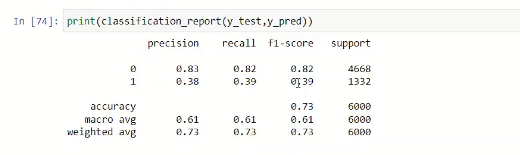
Accuracy for the decision tree is stored inside ‘a2’ and the results show that the decision tree is predicting the values with 72.6% accuracy and it is shown below. 27.4% is the mean\_absolute\_error value obtained for the decision tree.



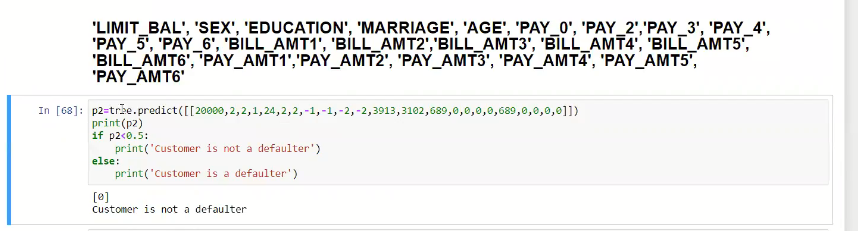
Cohen\_kappa\_score is identified for the decision and found 21.11% and in the other case, matthews\_corrcoef value is obtained as 18.63%.



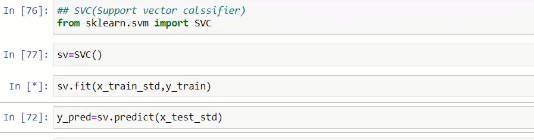
The following figure shows the classification report for the decision tree prediction for 0s and 1s individually.



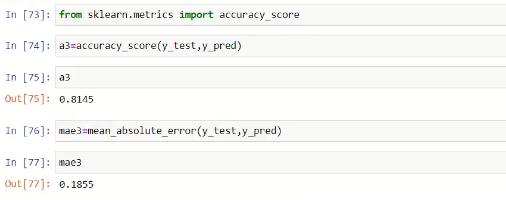
The same data is considered and given as input to the predict function and used as the tree variable in this case. The results show ‘1’ which means ‘Customer is a defaulter’



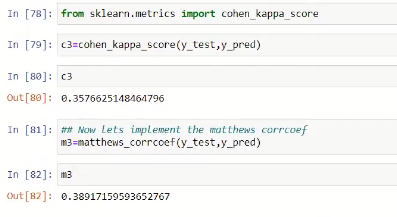
The next activity is implementing the Support Vector Algorithm (SVC) which is the third algorithm considered for the project and SVC is imported from sklearn.svm model. SVC function is stored in a variable sv and used for implementing fit function by passing training datasets such as x\_train\_std and y\_train. Then, y\_pred values are identified by passing input values such as x\_test\_std by using predict function and sv variable as shown below.



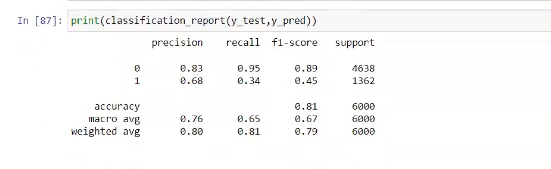
The results of the accuracy score and absolute errors are given in the figure below. 81.45% of accuracy is obtained for the SVM algorithm and 18.55 is the error rate.



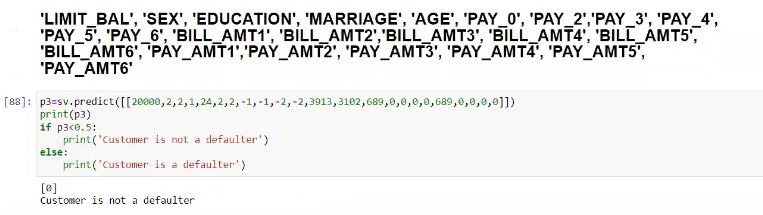
From the results, the cohen kappa score is obtained as 35.66% whereas 38.91% is the score obtained for matthews corrcoef.



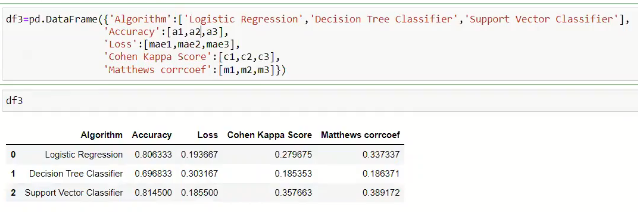
Following figure shows the classification report for the SVM algorithm.



Same as in previous cases, the same sample data is provided to predict function by using the SVM algorithm. The SVM results in the value of 0 and it indicates that ‘Customer is not a defaulter’



These are the results obtained for each of the algorithms and it is important to compare them all to find the best one. For that purpose, data frame df3 is defined and passed the variables of accuracy (a1,a2,a3), loss (mae1,mae2,mae3), cohen kappa score (c1,c2,c3) and matthews corrcoef (m1,m2,m3) variables above and the results are shown below.



The results show that Support Vector Classifier has the highest accuracy with 81.45% compared with the Logistic Regression and Decision Tree having values of 80.63% and 69.68% respectively. Also, the percentage of loss is very less for SVM than the other algorithms. Kappa Score and Matthews corrcoef are higher for SVM.

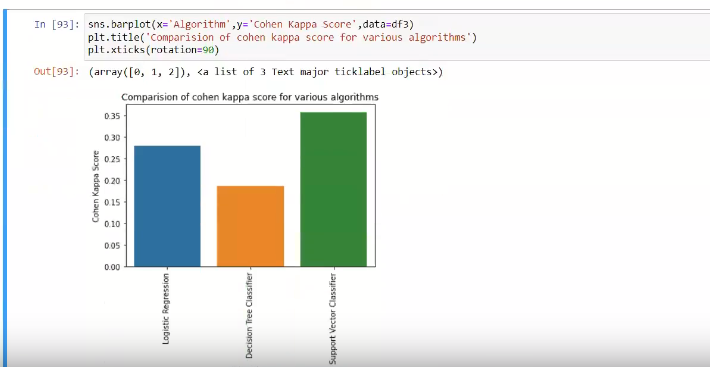
The next activity is creating a plot by considering algorithms on the x-axis and accuracy in the y axis and the plot is shown below.



Similarly, the plot is created to analyze the level of losses or errors for each algorithm.



The following figure demonstrates the cohen kappa scores for each algorithm.



Comparison is made on the matthews corrcoeaf scores obtained for each algorithms which are shown below.



The machine learning algorithms are successfully implemented, executed, and got the results without any errors. The quality dataset is used after ensuring no null values, checking data types, eliminating ID columns and separated input and output columns, and preprocessing using StandardScalar. The graphs and numbers indicate that SVM is giving a higher performance in terms of accuracy, precision, recall, and f1-score compared with decision tree and logistic regression. Also, the scores of cohen kappa and mathews corrcoef is good. These results indicate that SVM can be used to train a machine to make a good prediction in identifying the credit card defaulter. By using that, the machine can automatically make decisions based on train data and it simplifies the managing of credit card customers by the company.

# Discussion

The experiments result based on the machine-learning models are presented in Table 1 and Table 2. Totally, four performance measurements derived from the confusion matrix are used to evaluate the performance of each model implemented for predicting the credit card defaults. Those performance metrics include accuracy, precision, recall, and F1-score. In addition, there were other three important quality measures used in the study to determine the validity and reliability of the test and measure the quality of the classification. These measures include the cohen kappa score and Matthews correlation coefficient.

**Performance of the Machine Learning Models**

As shown in Table 2, the three machine-learning models were evaluated for the prediction performance. Among all models, the Decision-Tree model was inferior to the other two models in terms of accuracy metric, although the remaining two models achieved an accuracy score above 0.70. Because 70 percent of the collected defaults cases in the selected dataset were found as positive, the study dealt with an imbalanced dataset. From the experiment, it is observed that the Decision-Tree model obtained the lowest accuracy score.

Table 1: Precision, Recall, F1-Score of Each Classifier

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S No** | **Algorithm** | **Precision** | | **Recall** | | **F1-Score** | |
|  |  | ***Not Defaulter*** | ***Defaulter*** | ***Not Defaulter*** | ***Defaulter*** | ***Not Defaulter*** | ***Defaulter*** |
| **0** | Logistic Regression | 0.81 | 0.72 | 0.97 | 0.24 | 0.89 | 0.36 |
| **1** | Decision Tree Classifier | 0.82 | 0.36 | 0.78 | 0.42 | 0.80 | 0.38 |
| **2** | Support Vector Classifier | 0.83 | 0.68 | 0.95 | 0.34 | 0.89 | 0.45 |

Table 1 shows the accuracy, precision, recall, and f1-score of each model. The Logistic Regression Model was used to predict the credit card defaults. The experiment produced a classification report covering the test results of each machine learning model performed on the trained data. This report presented that the precision value of LR is obtained 0.806 (or 0.81), which indicates the precision of the test is 80.6 percent. Considering the model's performance, the precision of test results for a user is not a credit card defaulter is 81 percent. The model maintains the precision of 72 percent when it predicts if the user is a credit card defaulter. The recall value of the LR model is obtained as 0.97. The recall value found that true positives in the dataset are 97 percent correct. F1-score of the LR model is obtained as 0.89, which is a weighted average score of both precision and recall values.

In the binary classification report, the three performance measures scores (i.e., precision, recall, and f1-score) of the Decision Tree (DT) model are 0.82, 0.78, and 0.80, respectively. Comparing precision scores of both LR and DT models, it is found that the DT model has shown slightly higher precision than LR. But, recall and f1-scores are less than the LR. The Support Vector Classifier (SVC) model's performance scores were higher than both LR and DT models. The values of the three performance metrics of SCV were 0.83, 0.95, and 0.89, respectively. Therefore, the SVC model has generated better results than LT and DT models.

Table 2: Credit Card Defaults Prediction Results: Accuracy, Loss, Cohen Kappa, and Matthews Correlation Coefficient

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S No** | **Algorithm** | **Accuracy** | **Loss** | **Cohen Kappa Score** | **Matthews corrcoef** |
| **0** | Logistic Regression | 0.806333 | 0.193667 | 0.279675 | 0.337337 |
| **1** | Decision Tree | 0.696833 | 0.303167 | 0.185353 | 0.186371 |
| **2** | Support Vector Classifier | 0.814500 | 0.185500 | 0.357663 | 0.389172 |

The above table 2 shows that the Support Vector Classifier's accuracy score is higher than Logistic Regression and Decision Tree classifiers. From the experiment results, the accuracy score of the SVC is obtained as 0.814, and the other two classifier's accuracy scores were 0.806, and 0.696 respectively. The results show that the DT model's predictive accuracy was considerably worsened than the other two machine-learning models. The accuracy score of the LR model is slightly lower than the SVC model.

Figure 2: Accuracy Scores of Classifiers Figure 3: Error/Loss Rate in of Each Classifier

Figure 4: Cohen Kappa Scores Figure 5: Matthews Correlation Coefficient Scores

The above figures 2, 3, 4, and 5 illustrate the comparison of accuracy scores, loss or error rate, Cohen Kappa Scores, and Matthews Correlation Coefficient of each algorithm or model. Observing Figure 2, the SVC model has high accuracy compared to Logistic Regression and Decision Tree model. When the model's accuracy score is high, the loss of accuracy or error rate is low. Figure 3 shows that the loss in accuracy of SVC is low compared to LR and DT models. DT has a high loss rate compared to SVC and LR. Similarly, the loss rate of LR is slightly higher than SVC.

While measuring the classification's performance, the Cohen Kappa metric was used to measure the agreement between two raters, that is, whether a user is a credit card defaulter or not. This measure compares the machine learning model predictions with the dataset's manually established credit card defaulter rating information. Figure 4 shows the comparison of Cohen Kappa scores of three machine learning models. From the graph, it is found that the SVC model has a higher Cohen Kappa score than the other two machine learning models. SVC model ensured high reliability and validity in its prediction whether the user is a credit card defaulter or not. The Cohen Kappa score of SVC, LR, and DT models is 0.357, 0.279, and 0.185.

Figure 5 illustrates the comparison between all three models' Matthews Correlation Coefficient (MCC) metric. This metric is a more reliable statistical rate that generates a high score only if a model's prediction obtained good results in all four confusion matrix categories such as TP, FP, TN, and FN, corresponding to the positive and negative elements in the dataset. Comparing the MCC scores of each model found that the MCC has produced a high score for the SVC model and a very low score for the DT model. The SVC model's MCC score is 0.389, which is fair. The MCC scores of LR and DT models are obtained as 0.337 and 0.186, respectively.

From the project, I learned about identifying the credit card defaults used in the machine learning algorithms. I got some good general knowledge on the importance of finding the credit-card ruins and current problems and found that machine learning algorithms can help avoid the existing problems. I got an idea about the practical implementation of applying the machine algorithms to identify and predict credit card defaults. The project is completed as per the plans and meets all the objectives defined. I have collected information about the credit card attributes, factors that lead to credit defaults and found information on the application of machine learning algorithms. Logistic regression, decision tree, and support vector classifiers are used for the project to predict credit card defaults. The project's aim is met as it proved that algorithms could help machines train and forecast the values. It can support the bank to make proper decisions so that credit card defaults can be reduced. All the activities are completed as per the schedule, time frame, and within the budget. The experience and knowledge from the project are very useful, and they can be applied to future projects.

# Conclusion, Future Study, and Self-Reflection

**Conclusion**

The banking and the financial sector are required to maximise the profits by measuring and controlling that arises by issuing credit cards to the customers. At the same time, it is required for the banks to issue credit credits as one of the services they are required to provide to compete in the financial market. Moreover, the performance of the credit card portfolio also does affect the performance of the commercial banks increasing customer satisfaction and also revenue for the bank. Thus it is required for the banks to make proper decisions in regards to their credit card portfolio.

The implantation of the project helped to develop a model to predict credit card defaults. To issue credit cards and set by the credit limit it is required for the banks to distinguish between good and bad customers. The credit card is generally issued based on the credit score and behavioural score of the customers. It helps the banks towards guiding the decision making. But the use of the credit score does not help to entirely understand the different factors that lead to credit card defaults. Understanding the factors is very important for the banks as they are faces a lot of competition in retail banking as they are required to relax the credit approvals and the procedure to appraise the customers providing them credits.

The document helped towards understanding the factors that do influence credit card spending includes the credit card attributes, demographic factors, bank policies and the attitude of the customers towards money. It also helped to understand the factors that lead to credit card defaults that include the total credit card defaults to income ratio, size of the credit limit, proportional payment of expenses, inadequate awareness. The implementation of the data science considered various factors that can lead to default credit card payment considered various factors that including limit balance, sex of the customer, age, marriage, pay limits, amount bill etc.

The data science project was implemented with the help of the Python language. The use of python helped to apply the ML techniques of the use of the various types of algorithms that helped to build up the prediction model. The process involved downloading the data and importing the various libraries in Python. Pre-pressing and Cohen’s Kappa Scorewas used for understanding the reliability of the score. Different types of machine learning algorithms were applied to generate a predictive model. The different types of algorithms that were used included logistic regression, decision tree classifier and support vector machines. By running various types of algorithms do indicate the use of the support vector machine does provide the best performance among the use of these algorithms. The use of SVM to provide the best model that helps effective prediction for credit card default. This is because of the use of support vector machines to provide the best accuracy, precision, recall and f1 score compared with the other algorithms that are used. The use of the SVM model can help the banks and other financial institutions to help to manage the credit card portfolio and to help to reduce the credit card default risks.

**Future Study**

The future study holds a new direction for the use of machine learning to be used in the credit card industry. The future study of credit cards defaults is required to include various other factors that would help to enhance defaulter prediction. It is required to consider additional data that includes the current wealth, property ownership, insurance claims, the use of social network profiles and the various types of other personal indicators that would help to create a more holistic profile of the borrowers. The use of such an approach as the future study will help the bank and financial institutions to integrate the customers with having a weak credit score. In the credit card industry machine learning can also be used for effective targeting and retaining the customers and to ensure effective customer care. Future study is required to target the use of advanced machine learning techniques like the use of Natural Language Processing (NLP) application. Moreover, the future also holds the use of deep learning that would help to classify the credit card defaulters and towards detecting human behaviours. The use of various deep learning algorithms like CNN can help to increase the level of accuracy of the SVM prediction model.

**Self-Reflection**

The project provided me with a practical awareness of credit card default risks and the importance of migration strategies and the best decision making alternative that the banks are required to employ. The project also helped me to nurture research skills with the use of keywords and also my writing skills. The learnings helped me to nurture my skills in data science projects and the use of machine learning that would help to predict the future and help the bank classify the defaulters. The project helped to understand how machine learning can be used, the process that is required to be applied and the use of the different types of algorithms and assessing the best algorithm for prediction models. The learning also reflected that analyzing the risk promptly is very important for the banks and the use of conventional models do have certain limitations. The use of machine learning algorithms can help to enhance the level of predictions. The project helped me to enhance my experience of the use of machine learning tools and the way towards generating the model. The project excellently helped me to learn new things as an aspiring data scientist. It helped me to take the first step towards reaching my goals. The project also provided me with the experience to learn the techniques and the use of Python to do a data science project. It helped me to enhance my skills in Python coding. It also helped me to enhance my soft skills with the ability development to read and write the codes and understand the same. The project helped me to take the first step for my career development in the future.

# References

Ala’raj, M., Abbod, M. F. & Majdalawieh, M., 2021. Modelling customers credit card behaviour using bidirectional LSTM neural networks. *Journal of Big Data,* 8(69), pp. 1-27.

Banker, S., Dunfeld, D., Huang, A. & Prelec, D., 2021. Neural mechanisms of credit card spending. *Scientifc Reports,* 11(4070).

Butarua, F. et al., 2016. Risk and risk management in the credit card industry. *Journal of Banking and Finance,* Volume 72, pp. 218-239.

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

Delamaire, L., Abdou, H. & Pointon, J., 2009. Credit card fraud and detection techniques: a review. *Banks and Bank Systems,* 4(2), pp. 57-68.

Han, Y. et al., 2020. Detection and Analysis of Credit Card Application Fraud Using Machine Learning Algorithms. *Journal of Physics: Conference Series,* Volume 1693.

Hodson, R., Dwyer, R. & Neilson, L., 2016. Credit Card Blues: The Middle Class and the Hidden Costs of Easy Credit. *Sociol Q.,* 55(2), p. 315–340.

Inman, P., 2019. *Default rate on UK credit card debt at highest for two years.* [Online]   
Available at: https://www.theguardian.com/money/2019/apr/18/default-rate-uk-credit-card-high-banks  
[Accessed 30 10 2021].

Jantsch, L., Becker, J. L., Solana-González, P. & Vanti, A. A., 2021. Analysis of default risk in credit card use. *Brazilian Journal of Development,* 7(6), pp. 62634-62656.

Khare, A., Khare, A. & Singh, S., 2012. Factors affecting credit card use in India. *Asia Pacific Journal of Marketing and Logistics,* 24(2), pp. 236-256.

Kumar, S. & Chong, I., 2018. Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States. *International Journal of Environmental Research and Publc Health ,* Volume 15, pp. 1-24.

Lee, J. M., Park, N. & Heo, W., 2019. Importance of Subjective Financial Knowledge and Perceived Credit Score in Payday Loan Use. *Int. J. Financial Stud,* 7(53), pp. 1-21.

Manir, S. P., Saini, A., Sarkar, S. D. & Ahmed, S., 2019. Credit Card Fraud Detection using Machine Learning and Data Science. *International Journal of Engineering Research & Technology (IJERT),* 8(9), pp. 110-115.

Marsden, P., 2019. *Digital Quality Management in Construction.* s.l.:CRC Press.

Ma, Y., 2020. Prediction of Default Probability of Credit-Card Bills. *Open Journal of Business and Management,* Volume 8, pp. 231-244.

Muni, J.-H. & Jung, S. W., 2021. A customer credit Prediction Researched to Improve Credit Stability based on Artificial Intelligence. *Korean Journal of Artificial Intelligence,* 9(1), pp. 21-27.

Odhiambo, A. A. & Memba, F. S., 2012. Credit Cards and Performance of Commercial Bank Portfolios in Kenya. *International Journal of Arts and Commerce,* 1(6), pp. 167-173.

Pedro Larrañaga, D. A. J. D.-R., Ogbechie, A., Puerto-Santana, C. E. & ·, C. B., 2018. *Industrial Applications of Machine Learning.* s.l.:CRC Press.

Pumsirirat, A. & Yan, L., 2018. Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. *International Journal of Advanced Computer Science and Applications,* 9(1), pp. 18-25.

Shift, 2021. *Credit Card Statistics.* [Online]   
Available at: https://shiftprocessing.com/credit-card/  
[Accessed 1 11 2021].

Soll, J. B., Keeney, R. L. & Larrick, R. P., 2013. Consumer Misunderstanding of Credit Card Use, Payments, and Debt: Causes and Solutions. *Journal of Public Policy & Marketing,* 32(1), p. 66–81.

Teoh, W. M.-Y., Chong, S.-C. & Yong, S. M., 2013. Exploring the factors influencing credit card spending behavior among Malaysians. *International Journal of Bank Marketing ,* 31(6), pp. 481-500.

Teoh, W. M.-Y., Chong, S.-C. & Yong, S. M., 2013. Exploring the factors influencing credit card spending behavior among Malaysians. *International Journal of Bank Marketing,* 31(6), pp. 481-500.

Visa, S., Ramsay, B., Ralescu, A. & Knaap, E. v. d., 2011. *Confusion Matrix-based Feature Selection.* s.l., Conference: Proceedings of The 22nd Midwest Artificial Intelligence and Cognitive Science Conference.

Warrens, M. J., 2015. Five Ways to Look at Cohen’s Kappa. *Psychology & Psychotherapy,* 5(4), pp. 1-4.

Yaseen, M. K., Raheem, M. & Sivakumar, V., 2020. Credit Card Business in Malaysia: A Data Analytics Approach. *(IJACSA) International Journal of Advanced Computer Science and Applications,,* 11(12), pp. 383-390.

Yayar, R. & Karaca, S. S., 2010. Identifying the Factors Affecting the Consumer Credit Card Ownership: Empirical Evidence from Turkey. *Journal of Aplied Economic Sciences,* pp. 195-204.

**Appendix**

****