2020

A Model for predicting credit card loan defaulting using cardholder characteristics and account transaction activities

Muchiri, L. Munene
Faculty of Informatiopn Technology
Strathmore University

A Model for Predicting Credit Card Loan Defaulting using Cardholder Characteristics and Account Transaction Activities

By

Muchiri, Lewis Munene

113413

A Thesis submitted to the Faculty of Information in partial fulfillment of the requirements for the award of Master of Science in Information Technology

Strathmore University

April 2020

Declaration and Approval

I, Muchiri, Lewis Munene, declare that this research has not been submitted to any other University for the award of a Degree in Information Technology.

Student Name: Muchiri, Lewis Munene

Sign: _____ Date: _____

Supervisor's Name: Dr. Vincent Omwenga

Sign: _____ Date: _____

Abstract

Defaulting of credit card loans is an area of great concern amongst banks and financial institutions. This is because loan portfolio is considered an asset to the institution and one which directly impacts the firm's profitability, hence effective measures need to be put in place to ensure that default risk is at manageable levels. Credit card accounts are normally classified as "good" or "bad" depending on whether the cardholders are able to repay their debts within the agreed time or not. Good accounts continually settle their debts as per the agreement with the bank and consequently receive a higher credit limit over time allowing them have more credit at their disposal while bad accounts default their payments and end up with blocked credit cards or have more punitive measures taken on them. The latter bring losses to the banks. It is therefore imperative for these financial institutions to improve their credit management processes with the aim of optimizing their provisioning for bad debts, minimizing losses resulting from defaults and maximizing on revenue that would be generated from the "good" accounts. Traditional credit scoring techniques have relied on static models of loan default prediction which produce classes of cardholder accounts according to default risk but which give no insight to the changes in loan states over time. Changes in loan states may reveal much about the terminal state of the credit card loans which is to be expected with respect to the defined duration of time and was hence a good consideration for study. This study applied Logistic Regression Analysis to develop a model that learned patterns from observed account transaction activities and performed predictions on subsequent credit loan states based on the learned information. Successful predictions were obtained with an accuracy level of 85.28%, a recall level of 83.71% and a precision level of 77.87%, indicating the value in considering the states of loans over time and the events leading up to the terminal default or non-default states as opposed to a singular focus on the final default or non-default events.

Keywords: Logistic Regression Analysis, Account Transaction Activities, Cardholder Characteristics

Table of Contents

List of Tables

List of Figures

List of Equations

Chapter 1: Introduction

1.1 Background

A credit card is a financial instrument issued by banks to their customers allowing them to purchase goods or services on credit (Rajani, 2009). It is normally a branded plastic card having one or more of the card payment interfaces: magnetic stripe, contact chip and/or an embedded Near-Field Communication (NFC) chip. When customers use their credit cards, they are required to repay the credit extended to them plus any interest charges that may have been incurred over the duration of usage of the card.

In order to issue a credit card, a bank maintains and builds a credit score for a customer over time based on their account usage in terms of transactions history and how well the bank knows that customer, a metric derived from the bank's Know Your Customer (KYC) process. The customer then applies for a credit card and is issued one, together with a credit limit and terms of credit. Ideally a credit card begins with a balance of zero and increments upwards as the cardholder uses it to purchase goods or services (Shain, 2019). Consequent to this, the bank compiles information on the cardholder's purchases throughout the course of the set billing cycle (typically thirty days) and sends one bill at the end of that cycle, thereafter expecting a full or partial repayment within twenty-five days after the end of that billing cycle (Evans & Schmalensee, 2005). Cardholders who pay off their balance in full every month do not pay any interest, and this does not generate interest revenue for the card issuer (Johnson, 2011). Those who pay the minimum required amount per month repay the rest with interest, and late payments attract penalties on those accounts (Khan, 2011, p. 15). This describes the revolving credit principle, where a cardholder can carry forward their balance to the next repayment period and repay that balance with interest (Benson & Loftesness, 2014). Revolving credit cards will be the primary focus of this study.

The mechanism of credit card operation entails a cardholder performing purchase transactions at a merchant establishment using their card. This may be at a shop, website or a merchant payment gateway. The merchant transmits the cardholder's information and transaction details to their acquiring bank which then captures the transaction and routes it to the appropriate card network. The card network then routes the transaction to the cardholder's issuing bank for approval or decline (Dwyer, 2018). Following an approval, the merchant delivers the purchased goods or services and the issuing bank raises a bill on the cardholder. This bill has to be paid according to the agreement between the cardholder and the issuing bank (Gurusamy, 2009).

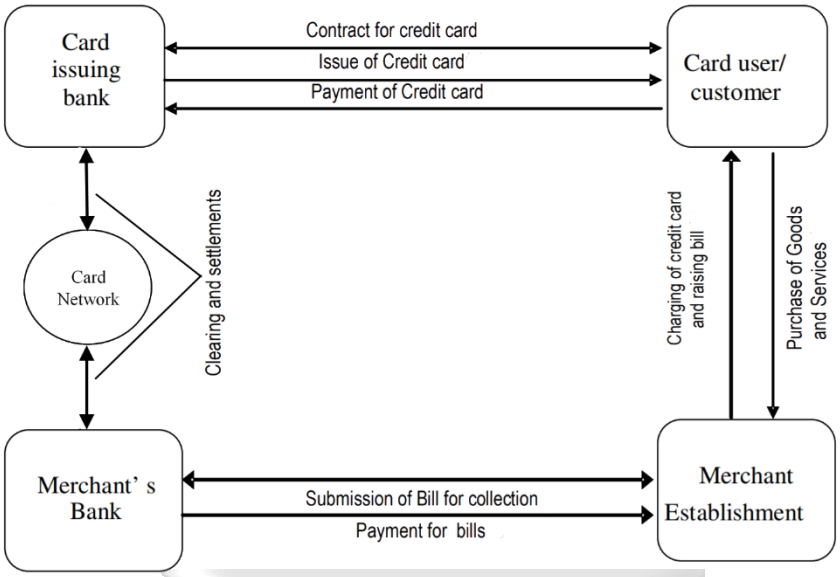

Figure 1.1: Mechanism of credit card operation (Adapted from Gurusamy, 2009))

When a late payment occurs, banks generally impose a late fee on that account, and upon prolonged defaulting of the loan, (normally 180 days), the bank writes off the loan and sells it off to a collection agency or lists the cardholder with a credit bureau as a defaulter (Mulligan, 2017). This adversely affects the customer's credit score with their bank and with other credit facilities. It is also a loss to the bank, hence it is of importance to implement measures that reduce the cases of loan defaulting by cardholders or the severity of loss caused by the defaults that occur. According to Islam (2018), reasons that cause people to miss payments for credit card bills include job loss, being steeped in excessive debt, inability to work and poor financial planning among others. These result in varying intensities of loan delinquency and influence the ability of customers to recover from delinquent states to favorable states, ultimately determining whether customers default or not.

Multiple studies have been undertaken to predict credit card loan defaulting using account transactions data, resulting in the development of various models for predicting credit card defaulting. In general, there are static and dynamic models of default prediction. Static models rely on classification of accounts as 'good' or 'bad' based on how the accounts perform over a fixed period of time. Bad accounts end in a defaulted state while good accounts end in a solvent state. Islam (2018), for example, achieved good results using a framework based on extremely randomized trees – an ensemble method that combines multiple decision trees to arrive at a classification. Dynamic models put more emphasis on the events leading up to a default state, and not on the final default state itself. This includes use of survival models to depict the

survivability of loans over time and probability of transition from one state to another. This work relies on the dynamic prediction models, and specifically considers Logistic Regression Analysis as a good fit for depicting the transitions that a loan goes through over time, putting into consideration that the state of a loan at a certain time, t, has a certain level of influence on the state that will be observed at time t + 1.

In order to guide the course of this study, a distinction between the terms "defaulting" and "delinquency" will be required so as to clarify the approach used. Defaulting is defined as the failure to pay the agreed amount as scheduled in the credit terms of agreement (Majaski, 2019). This is considered a terminal state, whereby accounts that enter this state do not leave it. Delinquency is the tendency to delay or miss regular instalment payments (Majaski, 2019) and may refer to any state in which the customer is not solvent enough to have settled the entire balance of their previous period's bill. This implies that delinquency has severity levels, while the default state may be viewed as the most severe level of delinquency that can be accepted by the bank.

1.2 Problem Statement

Loan defaulting by clients is undoubtedly one of the biggest issues faced by organizations in the financial industry and hence managing of firms' loan portfolios is top priority for the assurance of business continuity and sustainability (Philemon, 2018). This is because the loan portfolios of financial institutions have a direct impact on their profitability and therefore an improvement to the management of risk of loan defaulting can yield great revenue gains. On the other hand, poorly managed default risk leads to revenue losses. Financial institutions use different techniques to classify their customers as potentially good or potentially bad (Husejinovic, Keco & Masetic, 2018). These techniques rely on data that is available at time of credit card application to determine whether a customer qualifies for a credit card or not, and normally update their customers' credit ratings over time – either to give them a higher score or to block access to credit facility.

Over the years, lending institutions have relied on scoring models to distinguish bad borrowers from good borrowers and to determine the limits of amounts that can be advanced to specific borrowers (Oula, 2013, p. 3). While this is beneficial, greater value can be attained by complementing it with models that predict defaulting and repayment of loans as a way of forecasting and which can be used to inform the process of provisioning for bad debts. Studies done to predict loan defaulting have mostly emphasized on classification of accounts as 'good'

or ‘bad’ based on whether they eventually default the loan or not. This provides great value, however, there remains a lot of uncovered ground since with revolving credit, the lender is interested in identifying customers who do full repayment, those who do partial repayment and those who do not repay at all during any one billing cycle in all billing cycles that are considered before borrowers are declared as defaulters. This has great impact on the profitability of the organizations. Simply classifying the accounts as ‘good’ or ‘bad’ does not reveal the extent to which customers can be delinquent, and hence gives no hint as to how much a bank can actually make from its customers with respect to the credit card loans.

This research proposes a prediction model that considers the different states of loan delinquency that a cardholder may be in before finally reaching the terminal state of default and hence for any billing cycle, and given information on that customer’s performance up to that cycle, it may be possible to predict the performance of that account for the next billing cycle and possibly subsequent billing cycles up to the terminal period in which defaults are determined.

1.3 Aim

The aim of this research is to minimize the defaulting of credit card loans by cardholders by developing a model for predicting loan defaulting using cardholder characteristics and account transactional activities.

1.4 Specific Objectives

(i) To analyze the cardholder characteristics that influence defaulting of credit card loans
(ii) To analyze the account transaction activities that influence defaulting of credit card loans
(iii) To examine the methods and technologies that are used to predict defaulting of loans for credit card accounts
(iv) To develop a loan default prediction model for credit card accounts
(v) To test the performance of the model in predicting loan defaulting for credit card accounts

1.5 Research Questions

(i) How do the cardholder characteristics influence credit card loan defaulting?
(ii) How do the account transaction activities influence credit card loan defaulting?
(iii) What are the methods and technologies that are used to predict defaulting of credit card loans?
(iv) How can the loan default prediction model be developed and what algorithms and methodology will be used in its implementation?

(v) How can the performance of the developed model be tested?

1.6 Justification

This study was done to provide a means to improve the management of loan portfolios for issuers of credit cards through the ability to predict defaults on loans by analyzing cardholder characteristics and account transaction activities of the customers. The value of this outcome is that it aids in the decision making process that is concerned with which customers to issue loans to and how much of it to issue to them. Poor lending decisions lead to significant financial losses by institutions and therefore loan default risk has a major impact on the wellbeing and profitability of the financial institutions. As a result of this, approaches that aid the process of managing default risk are a key concern for these institutions and should be included among the priority issues that have to be addressed.

Whilst much has been done in the area of predicting credit card loan defaulting, Leow and Crook (2014) indicated that an alternative perspective in modeling credit risk exists in considering the distinction between defaulting and loan delinquency. Defaulting is a terminal state of a credit card loan, where the customer's portfolio has become so adverse that the bank deems them as defaulters and takes action against them. The time period given before a bank can consider their customer as a defaulter varies from bank to bank based on their risk portfolio and the terms of agreement between them and their clients. To enrich the default prediction models, Leow and Crook (2014) showed that there are advantages to be realized in predicting the different states of delinquency that lead up to default of a credit card loan. Such advantages include the ability to more intricately compute a bank's economic capital not only during the time periods when default events are normally assessed but also during any future time period, and also the ability to gain insights into the factors that affect the movement towards or away from loan defaulting by cardholders.

Performing the analysis of default severity over time - often based on months or the defined loan repayment cycles makes the problem a discrete-time-based prediction problem where it is possible to predict repayment rates of clients throughout the time periods before the default event is deemed to occur. This yields a solution that can show the delinquency rates of clients and can eventually predict the default event itself and which can help the banks to more effectively manage their credit risk and ultimately reduce the default rate for credit cards loans.

1.7 Scope and Limitation

The scope of this research is limited to learning from data related to credit card accounts and predicting the possibility of loan defaulting by analyzing the patterns exhibited within the data

This research used a publicly available online dataset containing credit card account information and the variables contained in that dataset. Obtaining such datasets directly from financial institutions proved difficult due to the sensitive nature of the data contained in them and the privacy implications that are present for the cardholders and for the institutions.

Chapter 2: Literature Review

2.1 Introduction

This chapter focuses on credit card loan defaulting and the studies explaining cardholders' behavior in their use of credit cards. It also includes reviewing the variables that most significantly impact the possibility of defaulting and the methods that have been used by banks to respond to delinquent accounts. Also, consideration is made on the various machine learning approaches that have been used to predict the possibility of defaulting of credit card payments and how credit accounts are classified by banks with regards to risk, and the different responses of the banks towards the accounts that fall into these classes.

2.2 Dimensions of Personal Characteristics of Credit Cardholders

Various characteristics affect the behavior and usage of credit cards by cardholders and in different measures. One way of distinguishing between these characteristics is by considering the attributes that describe the nature and state of a cardholder and those which describe the transaction activities between the cardholder and the bank (Lee, Lin & Chen, 2011). The former characteristics are henceforth regarded as Personal Characteristics of the cardholder and can collectively be defined as the factors which constitute the socio-demographic and socio-economic profile of the cardholder. Socio-demographic factors capture the interaction and culture of everyday life and the characteristics that are used to describe a population and include attributes such as age, marital status, gender and education level of people (Kiarie, Nzuki & Gichuhi, 2015). Socio-economic factors comprise the characteristics which directly influence the economic performance of a cardholder and include income, property ownership and occupation (Kanyi, 2009).

A solid understanding of how cardholders' personal factors influence the overdue risk of credit card loans is key for banking executives to make decisions regarding credit standards, loan amounts and criteria assessment in order to establish a risk control mechanism (Lee et al., 2011). For banks, income obtained from credit cards has significantly shifted from transaction fees towards the revolving interest rate income of credit card loans, and great emphasis is therefore to be put on maximizing the revenue generated from interest rates on these loans, which implies making a careful selection of the customers and the limits to set for them (Lee et al., 2011). Analyzing the personal profiles of these cardholders is a step towards achieving this objective.

In consideration of the personal characteristics, Gan, Maysami and Koh (2008) stated that cardholders with higher occupation level have higher stability of work and their risk of defaulting credit card loans is lower. This is because such people change work less often and hence the steadiness of monthly income for them is assured. This reduces their chances of defaulting on their credit card loans and makes it easier to reach them and follow them. The overdue amount for their loans is also relatively lower hence the higher the cardholder's occupation level, the lower the risk of defaulting. Lee et al. (2011) confirmed this hypothesis in their study and concluded that customers having no fixed work structure or with less formal and more flexible occupation levels have higher credit risk than those with higher and more rigid work structures. In particular, the study pointed out that military staff and civil service workers had lower overdue amount for credit cards, while workers in agriculture or fishing had higher overdue amounts. It was hence possible to categorize customers in terms of risk based on their occupation level.

Other explanations for cardholder behavior look towards consumer psychology and marketing instincts – with focus on factors such as self-control against impulsive buying. Various views exist, and notable is the view that the liquidity provided by credit cards makes it more likely that a consumer will purchase a given product, and would even increase the amount that the consumer would be willing to pay for the item (Bertaut & Haliassos, 2006). This may also explain why merchants are willing to provide the credit card as a payment method despite the transaction charges that are associated with it.

The idea that cardholders' personal characteristics are a driving force that influence, and to some extent determine the behavior of a customer with respect to their credit card usage is a key consideration to be held in studies aimed at predicting the loan status of customers after certain specified periods. This forms a solid foundation for their inclusion as study variables in this research.

2.3 Account Transaction Activities Influencing Defaulting of Loans

The main reason for having a credit card is to facilitate payments (Lee et al., 2011). Credit cards allow banks to extend credit to their customers with contract terms directing the transaction activities that occur on the credit card accounts, and covering issues such as the maximum credit amount that can be availed to customers, repayment cycles, interest fees and other charges incurred in the possession and usage of the credit cards.

As defined by Kocenda and Vojtek (2009), behavioral factors are the indicators of a customer's behavior with regards to their usage of credit cards. These are reflected by the account transaction activities that are undertaken by an accountholder in the course of utilization of the cards issued to them by their banks. The way customers utilize their cards is expressed through the transactions that they undertake and this is in light of parameters that directly constrain or guide how cardholders transact with their cards. These parameters are determined before loans are issued and may be altered according to the behavior of the customer in the duration in which their loan is active. An example of such parameters is a customer's credit limit, which is given an initial value by the bank, but is altered over time depending on the customer's uptake of loans and their repayments. Credit limit however, is only a reflection of the measure of creditworthiness that a bank has assigned to their customer. Marjo (2010) includes both the transaction activities themselves and the restraining variables that characterize a cardholder's relationship with their bank as part of the behavioral variables, even if they are not in themselves the transaction activities, but are directly adjusted based on the transaction activities. Consequently, it is acceptable to state that behavioral variables include factors such as credit limit (loan size), interest rate for individual cardholders, card transactions done by the cardholder and purpose of loan.

Account transaction activities include the following:

i. Purchase transactions (spending)

   This entails processing of credit card sales at a merchant point-of-sale device or card gateway. The card information is sent to the card transactions processor which sends the transaction to the card network and which then routes the transaction to the cardholder's issuing bank for approval of the transaction. An approval code is returned to the merchant if the transaction is approved, and the customer receives the goods or services paid for (Reid, 2019).

   For credit cards, a purchase activity is synonymous with borrowing since credit cards are an extension of credit such that a cardholder uses money that they do not have in their account at the moment, but with an intention to repay at a future date. At the end of the billing period (typically a month), all purchase transactions are aggregated to give the total bill amount that is owed by the customer to the bank. This becomes the customer's debt to the bank for that period.

ii. Pre-authorization and Capture transactions

Pre-authorization is similar to purchase with the exception that it does not complete the sale and funds are not debited from the card, but are instead held for between 7 to 10 days. A "Capture" request will then later be sent by a merchant to complete the initial sale request. This method is suitable for hotels, gas stations and the like, where a merchant needs to be sure that a certain amount is available in the card account before offering the service (Reid, 2019).

The net effect of a pre-authorization and capture transaction is similar to the effect of a purchase transaction in that the customer's total bill is increased. The difference between the two is in the point of execution of the debit activity, where funds are channeled from a customer's account to a merchant's account.

iii. Refunds

This performs the reverse of a purchase transaction and money is credited back to the customer's card. Transaction fees are also charged to the card when a refund is performed (Reid, 2019).

This kind of transaction activity reduces the customer's bill amount by the value of the original purchase transaction but also increases the bill amount by the refund fee that is charged. It is not expected that the fee be more than the reversed amount but nevertheless an increment to the customer's monthly bill amount is done.

iv. Verify

This is a zero-amount transaction that is performed to verify the card's validity (Reid, 2019). It is usually done on 'card-not-present' points of payment, such as e-commerce websites and typically works under the assumption that if the issuing bank approves the 'verify' transaction then the card is a valid and operational card. An account having more of these 'verify' transactions shows that the card is frequently being used and on many different e-commerce systems to perform purchase transactions. This could increase chances of default since the card is being heavily relied upon to perform purchases.

v. Balance Repayment Transactions

This refers to paying off the credit card debt. This can be done through multiple methods, as provided by the customer's bank. Some of the payment options include cash payment made to a bank teller, payment at an ATM (Borchetia, 2018), through mobile money or other means provided by the bank.

Repayment is normally done according to the regulations specified by the bank with regards to deadlines for repayment, minimum required repayment amount, and determines the delinquency level of the cardholder.

With regards to the account transactional activities, various patterns have been seen to impact the default level customers and thus influence the credit risk for those cardholders. Lee et al. (2011) stated that a higher ratio of the used credit to the credit card limit increases the possibility that the credit card loan will be overdue, meaning that higher utilization rates of credit cards increase the chances of delinquency. The credit limit sets the constraints for the cardholder's balance, and consequently the debt amount and indicates the amount of risk that a bank is willing to tolerate (Gross & Souleles, 2002). The more the available limit is used (with respect to the limit itself, that is, ratio of used credit to the limit), the more delinquent the cardholder will be and hence the more likely to default that cardholder will be.

2.3.1 Types of Credit Cards and their Transaction Activities

Transaction activities also vary based on different types of credit cards that are available. Credit cards come in various packages and products targeted at different kinds of cardholders based on their spending power and business needs. All this is with the aim of diversifying a bank's profit-making methods through use of credit cards, but all operating under the basic principle that the credit card is a financial instrument through which issuers can extend a line of credit to their cardholders and receive repayment for the credit at the contract-defined durations and schedules. Below is a non-exhaustive list of credit card types and the transaction activities associated with them:

i. Revolving credit cards

   These types of cards allow the cardholder to borrow money and then pay a minimum predetermined amount each month while carrying forward the balance to the next repayment period and paying an interest on the balance. The cardholder may also pay off the whole amount during the current month and pay no interest for that period (Benson & Loftesness, 2014).

ii. Charge cards

   In contrast to the standard revolving credit cards, charge cards do not allow customers to carry forward their balance, but instead the balance must be paid off by the set date in that period. Delays in payment attract fines on those accounts (Benson & Loftesness, 2014).

iii. Premium cards

   These are a type of brand that includes gold, platinum and other card products (as defined by the card network) and come together with premium services and benefits such as travel insurance and access to certain events, travel packages and other forms of insurance. In addition, an annual fee is paid for the card (Steele, 2019).

iv. Affinity and cobranded cards

These cards are associated with institutions to which they are issued. Affinity cards carry the name and brand of the institution, for example, a school or club while co-branded cards are sponsored by multiple parties, one being the issuer and the other being the retail institution (Benson & Loftesness, 2014).

v. Corporate cards

Corporate cards are issued to employees of certain organizations for limited use such as travel and entertainment purposes (Benson & Loftesness, 2014).

2.3.2 Payment Transaction Processing

Processing of transactions in payment card networks is facilitated by a payment infrastructure which is made up of computer networks, point of sale terminals, software applications and computers of financial institutions. Customers' financial data is contained in smart cards (chip cards, contactless Near Field Communication (NFC) cards), digital tokens, magnetic stripe cads and other tags made into wearable or attachable forms such as wrist-bands, stickers or key-chains. All these together constitute the framework for the electronic processing of card payment transactions (Radu, 2003, p. 9).

According to Radu (2003), the roles/parties involved in payment transaction processing include the issuing bank, the cardholder, the merchant, the acquiring bank, the card association and the settlement bank. The issuing bank is a licensed financial institution that issues credit cards to its customers and that authorizes payment transactions. The cardholder is the customer of the issuing bank who uses credit cards to make purchases at a merchant store. The merchant is the party that accepts a credit card and provides goods or services to the cardholder. The acquiring bank is a licensed financial institution that provides banking services to a merchant and that acquires payment messages related to the payment transactions. The card association owns the card product and is responsible for exchanging messages between issuing banks and acquiring banks. The settlement bank is responsible for transferring funds between issuer and acquirer accounts based on the card transactions performed (Radu, 2003, p. 15).

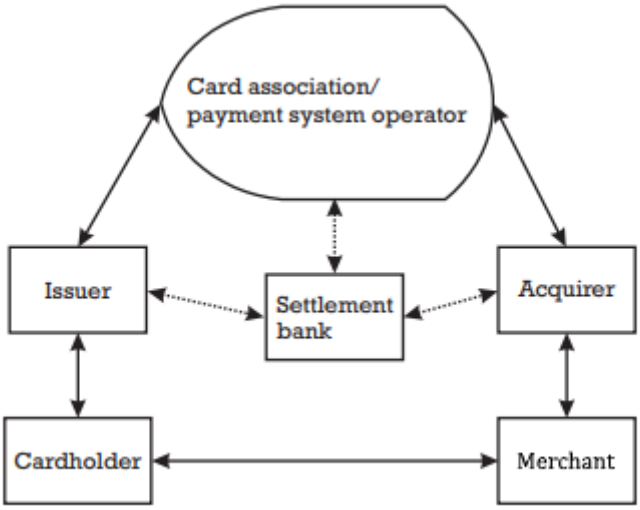

Figure 2.1: Parties Involved in Payment Transaction Processing (Adapted from Radu, 2003)

A payment transaction begins with a cardholder (the customer) presenting their credit card at a point of sale terminal or other reader device to make a payment. The transaction request together with the required financial details are sent to the merchant's acquiring bank via the network. If the customer's card was issued by the merchant's acquiring bank, then the bank authorizes the transaction (performs checks on the credit account and the merchant account and approves the transaction if all checks pass successfully). This describes an on-us transaction, where the merchant's acquiring bank is the same bank that issued the cardholder's credit card.

For off-us transactions, the acquiring bank sends a financial request to the card issuing bank through the payment network (VISA, MasterCard or any other card association) requesting a guarantee that the cardholder's account has sufficient funds to make the purchase, and also requesting confirmation. Once the confirmation is received, the transaction in authorized and a payment receipt is printed indicating that the payment was processed successfully. The acquiring bank receives payment from the issuing bank later through the settlement process, or according to the transaction rules at play within the card network in use (Radu, 2003, p. 12).

2.4 Empirical Review

The empirical literature review for this study was based on various notable studies undertaken to establish a link between cardholders' personal characteristics (socio-demographic and socio-economic factors) and behavioral and transactional attributes and the defaulting of credit card loans and to show the most powerful indicators of credit card loan delinquency over the lifespan of cardholder loans. This review also provides an understanding of the factors that can be relied upon to provide the most valid predictions for default and delinquency of credit card loans.

2.4.1 Cardholders' Personal Characteristics and the Tendency to Default

Personal characteristics of cardholders include the socio-demographic and socio-economic factors which both characterize the customer and his features at the time of application for a credit card (Marjo, 2010). These variables help to explain a customer's behavior with their credit card as they provide the lifestyle configurations under which the credit cards will be used. Highlighted in this study are the age, gender, education level and marital status variables, all of which have been shown empirically to have a predictive power for credit card delinquency.

The 'Age' variable defines a customer's age in years and is considered a categorical variable with bands to cover various age groups of individuals. Marjo (2010) defined 6 distinct groups, ranging from 0 (under 20) to 5 (over 70) and observed, in conformance with previous studies on default, that the tendency to default decreases as age increases, with the greatest risk being posed by 20-25 year olds. Marjo (2010) found that almost 45% of people in this group will end up defaulting on their payments which was explained by the assertion that older individuals are more risk averse, and are hence less likely to default on their loans while younger individuals take more risks. It is especially worse if the individuals are not economically stable, having not established themselves in their careers. The study done by Marjo (2010) did not include any observations for individuals below 20 years old, since the company policy (for the company under study) was not to grant loans to customers below 20 years of age. Similarly, in Kenya the minimum age for an individual to be recognized as an adult is 18 years, and hence it is extremely rare to find credit card packages for customers below that age.

Gender is a socio-demographic variable that is frequently used to predict credit default outcomes with consideration on the distinction between male and female cardholders. Marjo (2010) found that women default less frequently on loans as compared to men and Abdul-Muhmin &Umar (2007) also found that men have a higher tendency to revolve credit as compared to women, including in instances where interest changes continually. Further research done by Kanyi (2009) revealed that gender, if taken alone, does not influence credit card default rate. This was done using a Chi-square test for independence, which yielded results that showed that there was no statistically significant relationship between gender and default rate. This however was in contrast to former findings of Kocenda and Vojtek, (2009), which found that gender is indeed a risk factor in loans.

Education level refers to the highest academic qualification attained by a customer. In accordance with the findings of Kocenda and Vojtek, (2009), education level is a significant predictor of credit card default. People who have attained university degrees are shown to be less likely to default as compared to people who have only attained primary school or high school education, with the default rate being highest for people with only primary school education. Kanyi (2009) found that 16.8% of cardholders who have attained university level of education had defaulted as compared to 24.2% for those with a lower education level than the university education. There is therefore an inverse relationship between default rate and education level. The higher the education level, the lower the credit loan default rate for cardholders.

The study by Agarwal, Chomsisengphet and Liu (2010) led to the conclusion that a customer's marital status plays a statistically significant role in determining loan default outcomes. In particular, the finding was that a customer who is married is 24% less likely to default on their credit card debt as compared to one who is not married. The study suggested that married customers have a larger household income and can share the risk across members of the household, which hence explains the observation.

2.4.2 Cardholders' Transactional Activities and the Tendency to Default

The relationship between credit cardholders and their banks is expressed through the transactions that the cardholders perform with their credit cards. Transactional activities are both explanatory and response variables since while their general trends can be explained by other personal characteristics, they can still be used to show whether customers have delinquency tendencies based on observed patterns of occurrences. For example, as Lee et al. (2011) states, a higher ratio of used credit to the credit limit points to a higher tendency of delinquency and eventual default of loans. In this sense, they are useful predictors of delinquency rates amongst credit cardholders.

Attributes of the transactional activities include end of month bill amounts, repayment amounts and repayment dates (to indicate payment delays) for the bill statement balances. These attributes give a good picture of the cardholder's financial position and can be observed to trace the path leading to default of credit card loans. Customer's bill amounts are controlled by the credit limit set upon the customer by the bank since the customer cannot have a balance that is larger than the set credit limit for any monthly period. The credit limit may be altered over time, depending on the customer's behavior – for example it may be decreased when the

customer delays or pays less than the minimum required amount and may be increased when the customer exhibits good repayment tendencies. All these apply in conformance to the terms and conditions agreed upon between the customer and the bank.

A study done by Lee et al. (2011), revealed interesting findings of the transaction relationship between cardholders and banks. The first finding was that credit cardholders who pay a revolving interest rate have lower overdue loan amounts (lower delinquency levels) as compared to those who do not pay a revolving interest rate. The latter primarily pay off their debts in full without rolling forward the loan, but it was noted that they are indeed the ones who have higher overdue amounts when they can't pay their loans. The explanation given in that study for this observation was that even though the cardholders who roll forward their debts do not pay their amounts in full, they show responsibility for their debts since they at least choose to pay their minimum debts. A second observation was that customers with higher revolving credit amounts have higher overdue loan amounts. This means that the tendency to default increases as the loan amounts rolled forward increases. This can be checked as a ratio against the customer's credit limit – since all customers have their own limit amount, and the conclusion made that customers with higher credit amounts to credit limit ratio have a higher tendency to delay in their payments and eventually default on those payments. This is consistent with the next observation in the study, that the higher the ratio of used credit to the customer's credit limit, the higher the overdue amounts will be.

Transaction activities can be used as a powerful predictor of delinquency level by observing their trends with respect to the personal characteristics of the customers under scrutiny. They therefore formed a critical component in modeling delinquency rates and default risk as they are a reflection of a customer's ability to repay in time (when considered in retrospect) and also act as target variables to be predicted in future (when considered in prospect).

2.4.3 Methods and Technologies Used for Predicting Defaults

Every party faced with the decision to issue a loan or not to a beneficiary has to encounter the challenge of estimating whether the individual will honor the promise to repay or will default on the repayment either in full or partially. The desirable outcome is that the counterparty will repay (with interest for those whose terms dictate so) and in the agreed time. In order to make such an assessment, the lender considers how well they know the recipient based on their ability to repay, or the recipient's past records at repayment of loans, or the track record of similar borrowers during similar circumstances. What follows is the decision to lend, which carries

along the risk of loss in the event that the counterparty fails to fulfil the agreed terms. This is the basis for default prediction even in automated (computerized) environments (Brown & Moles, 2014).

Assessment of credit risk requires modeling of the probability of default for a counterparty. This can be used to determine whether to extend credit initially or to adjust the extended credit amount up or down as in the case of credit card loans issued to cardholders. The idea is to balance the gain from extending the credit (through interests) against the potential loss that would be incurred. Given that the probability of default for a certain repayment cycle is *p*, the following diagram depicts the assessment decision for the credit problem.

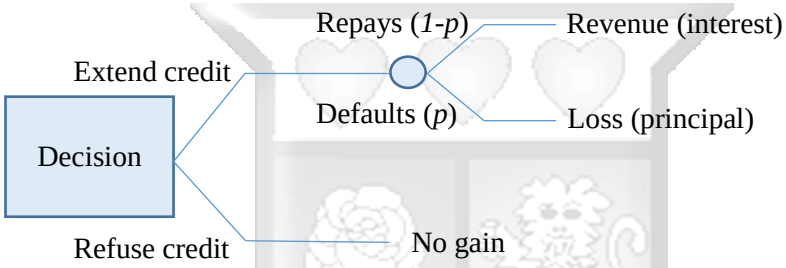

Figure 3.2: Assessment of Credit Risk (Adapted from (Brown & Moles, 2014))

The probability of default (*p*), is one of the key determinants for decision made as to whether to extend credit or to refuse the credit. That probability, as discussed in earlier sections, is itself determined by factors characterizing the loan recipient and can also be gleaned from the transactions that the customer has been performing over time. Therefore, the personal characteristics and transaction activities of the cardholders can be interrogated to give a good estimate for the value of the probability *p*, and hence form a basis for making the credit decision.

To extend this idea, and based on the problem at hand, the number of possible events resulting from the decision to extend credit to a cardholder may be more than two. For example, the revolving credit principle allows the bank to extend credit to a customer and receive a full or partial repayment at the end of the set period for repayment (Benson & Loftesness, 2014). If a partial payment is done by the customer, then the balance is rolled forward to the next repayment period (typically a month) and is expected to be repaid with interest during the next period. This necessitates that the bank imposes a minimum repayment amount upon the cardholder. If the cardholder repays in full, then no amount is rolled forward and no interest is

earned by the bank. Payments that are less than the minimum attract fines upon the account and cause a reduction in the amount of credit that can be extended to the customer on future periods. This scenario describes a multi-level view of loan repayment states, and assuming that *p* is the probability that a customer will not pay any amount, then the probability *1-p* is now divided amongst the remaining possibilities of loan repayment states as *a(1-p)*, *b(1-p)* and *c(1-p)* where $a + b + c = 1$. The following diagram depicts this multi-state scenario.

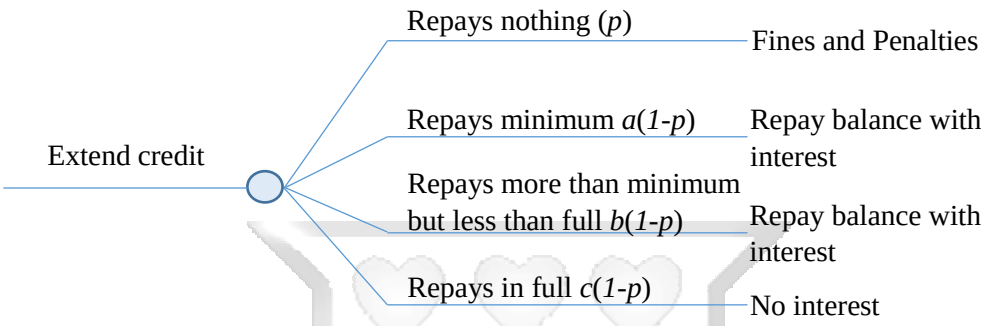

Figure 4.3: Assessment of Credit Risk (multiple states) (Adapted from (Brown & Moles, 2014))

Various approaches have been used to model the problem of credit default risk and have resulted in different outcomes for application in advising the risk management process of banks. Among these are classification approaches whose outputs include instance classes labeled as defaulters or non-defaulters, or labeled using the various risk levels that there are: low risk, medium risk, high risk etc. such include the work of Islam (2018) who used extremely randomized trees. Other approaches, such as the one employed by Leow and Crook (2014) considers the various levels of delinquency through which a customer would go over time before being deemed a defaulter. This was an intensity model based on transition probabilities which also resulted in the ability to do month-by-month predictions of default possibility for customer accounts.

The approach proposed in this study considers that credit cards operate on the revolving credit principle, which presents a unique way to tackle the problem and develop a means for prediction of defaults. The ability to roll forward a debt to the next billing period indicates a form of dependency of the events occurring during the different billing cycles, such that one the outcome of the loan state of one billing cycle is influenced by the outcome of the previous billing cycle. With this regard, a default event may be foreseen by analyzing the events preceding it, since the events are not independent of each other. One way of performing this analysis is by considering a regression approach, which models a relationship between

predictor and response variables such that unit changes in one or more of the predictor variable values causes an observed change in the response variable. Credit card loans may be modeled using discrete time events where event outcomes are determined upon the conclusion of a billing cycle. The default event may be viewed as a terminal state along the chain of events over the lifespan of the credit card loans for a single time. During that entire time duration, the customer may have traversed more than one loan states, some of which may be solvent and others delinquent states.

2.4.3.1 Linear Regression Analysis

This form of regression analysis establishes a linear model between a dependent variable, Y, and one or more independent variables X_1, $X_2 \ldots X_n$, where in the case of credit risk modeling, Y is the observed status of the loan at the end of some predetermined period while $X_1 \ldots X_n$ are the covariates describing an account and which influence the outcome of Y during the period of observation (Pershad, 2000). Since linear regression results in a dependent variable Y whose value is continuous in nature, for purposes of classification, a cut-off value has to be determined which is used to distinguish the accounts which will default from those which will not default. The mathematical representation of linear regression is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Equation 2.1: Linear Regression (Adapted from Meyer & Pifer, 1970)

This method was employed by Meyer and Pifer (1970) to predict bank failures given four groups of explanatory factors: local economic conditions, general economic conditions, quality of management and employee integrity. The researchers tested with different selections of the cut-off value in order to determine the best cut-off value for the independent variable for determination of failure. The conclusion was that one could correctly classify at least 80% of the sample and give prediction of future failures with about the same accuracy.

This method is not popular with studies involving loan default prediction and account classification since the problem demands that the outcome variable be a categorical variable (default or non-default), a scenario which violates the linearity assumption in normal regression. The preferred approach in many studies has been use of logistic regression which logarithmically transforms the output variable and makes it possible to have binary values which are desired in such problems (Agbemava, Nyarko, Adade & Bediako, 2016).

2.4.3.2 Logistic Regression Analysis

Agbemava et al. (2016) developed a model using logistic regression to determine the factors that were statistically significant in influencing loan defaults by customers in the microfinance sector in Ghana. The researcher utilized binomial regression in predicting default, with two levels characterizing the customers as defaulters or non-defaulters. The findings showed that of all the predictor variables available for the study, only six were statistically significant: marital status, number of dependents, type of collateral, customer's assessment, loan duration and loan type, and with these it was possible to obtain a predicted default rate of 86.67% for the microfinance institution's loan customers in Ghana. Logistic regression analysis was appropriate for that study since the nature of the problem demanded a categorical dependent variable and predictor variables present were also categorical variables.

Logistic regression is a mathematical modeling technique that describes the relationship between one or more independent variables and one categorical dependent variables through a linear function whose output depicts the probability that the instance belongs to one of the classes of the dependent variable (Agbemava et al., 2016). If the dependent variable has only two possible options, then it is said to be a dichotomous or binomial dependent variable, hence the analysis approach becomes a binomial logistic regression. Dependent variables with more than two levels are polytonomous or multinomial dependent variables, and the analysis approach used becomes the multinomial logistic regression.

The mathematical representation of Logistic Regression is as follows:

$$Y = \frac{e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}{1+e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Equation 2.2: Logistic Regression Equation (Adapted from Lando, 2004)

Where *Y* is the probability of a customer belonging to either of the groups (i.e. defaulter or not). For example, Y = P (1| x_1, $x_{2, \ldots,} x_n$) where the customer is a defaulter and Y = P (0| x_1, $x_{2, \ldots,} x_n$) where the customer is not a defaulter. This assumes that the values (0, 1) represent defaulter and non-defaulter respectively.

The formula described above results in a sigmoid curve when values are plotted on a two-dimensional axis, and whose inflexion point denotes the probability at which a customer jumps from one class and into the other.

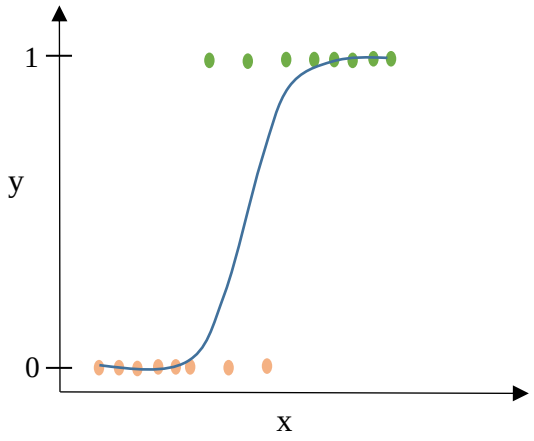

Figure 5.4: Logistic Regression S-Curve

To transform the logistic regression model into a linear one, the Logit function is used where logit is defined as the logarithm of the odds of a default event occurring (Lando, 2004).

Odds of default are defined as follows:

$$Odds(Y = 1) = \frac{p}{1-p}$$, where p is the probability of defaulting.

The log of odds (logit) is hence defined as follows:

$$logit(P(Y = 1 \mid x_1, x_2, \dots, x_n)) = log_e\left(\frac{p}{1-p}\right)$$, where p is the probability of defaulting.

The logit function transforms the logistic regression model into a linear model which takes the form

$$logit(P(Y = 1 \mid x_1, x_2, \dots, x_n)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Equation 2.3: Transformed Logistic Regression Equation (Adapted from Lando, 2004)

With this model, it is now possible to obtain a prediction in a two-level outcome situation (binomial logistic regression) or a multi-level outcome situation (multinomial logistic regression) by plugging in the values of the independent variables (X-variables) into the equation and performing a mathematical summation of all the values, then performing the inverse-logit of the result. The aim of logistic regression is to obtain the constant α and the coefficients of β_1 ... β_n of the independent variables.

In the case of Multinomial Logistic Regression, multiple models are generated (k-1 models, where k is the number of levels of the dependent variable). Here a reference level is selected, and used as the subject of the equation in which calculations for all other levels are included

(Williams, 2019). For example, for a dependent variable with three levels, A, B and C, and assuming that C is the reference level, then the formula for logistic regression will be:

$$P(C) = \frac{1}{1 + \left(e^{\left(\alpha^A + {\beta_1}^A X_1 + \cdots + {\beta_n}^A X_n\right)}\right) + \left(e^{\left(\alpha^B + {\beta_1}^B X_1 + \cdots + {\beta_n}^B X_n\right)}\right)}$$

Equation 2.4: Multinomial Logistic Regression Equation (Adapted from Williams, 2019)

Two models will be created, one which considers the log of odds of A occurring with reference to C, where C will be 1 – P(A) and B is not considered, and the other which considers the log of odds of B occurring with reference to C, where C will be 1 – P(B) and A is not considered. The instance will be assigned the level with the highest probability.

Multinomial logistic regression is appropriate for this study since both categorical and continuous independent variables are to be used to predict default levels, and also because the default levels are more than two (no repayment, minimum repayment, below full repayment and full repayment). In addition, multinomial logistic regression allows modeling of non-linear relationships (continuous and categorical variables to categorical outcomes) in a linear way through a logarithmic transformation which upon obtaining its inverse, will give a value indicating to which category the outcome variable will lie for each instance of the study.

2.4.3.3 Discriminant Analysis

Discriminant analysis assumes two groups of instances, one which is made up of cardholders who default on their loans over a period of time and the other in which the cardholders do not default over the same time period. Defaulters and non-defaulters are initially given, and the task is to determine the characteristics of cardholders who normally default and distinguish them from those who do not default. Using these characteristics, it is then possible to determine the class for a new cardholder, that is, whether they will default eventually or not (Lando, 2004).

Discriminant analysis possesses some weaknesses, which includes the assumption that instance characteristics follow a normal distribution, an assumption which is unrealistic for many observations that are made in practice. Also, the model is static and does not give information about how long a cardholder stays non-delinquent, which would also be valuable information for the financial institution issuing the cards (Lando, 2004).

2.4.3.4 Intensity Models and Transition Probabilities

Loan default risk models can factor in both conditional and unconditional probability of events in deriving models for default prediction. According to Holton (2013), a survival function *s(t)* is defined to indicate the probability of a loan surviving without default up to time t. Time divisions may be based on the bank's billing cycle or other form of division in which a default event may be determined for the loan. If the events observed at each time division operate by unconditional probability, then the probability of the loan being defaulted in time t+1 is given by *s(t) – s(t+1)*. For conditional probability, Bayes theorem is incorporated to yield the formula

$$\mathrm{Y} = \frac{\mathrm{s(t)} - \mathrm{s(t+1)}}{\mathrm{s(t)}}$$

Equation 2.5: Survival Function (Conditional Probability) (Adapted from Holton, 2013)

The survival function defines a mortality model for credit card default using discrete time and is derived by conducting a historical analysis of the account transactions data.

Switching the consideration from discrete time to arbitrary time intervals yields another form of loan default risk modeling which is based on expressing default risk as an average rate of default over a certain time period. Consider an arbitrary time interval Δt and a time t which is, say, three months. Assume that *s(3)* and *s(6)* are given. The conditional probability of default can be calculated between months 3 and 6 and an average rate of defaults per months can be obtained by dividing the conditional probability by 3. To get an instantaneous rate of default at any time t, take the limit as *t* goes to 0 using the following formula

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \,|T > t)}{\Delta t}$$

Equation 2.6: Hazard Function (Default Intensity) (Adapted from Holton, 2013)

This defines a hazard function *h(t)*, also called default intensity.

Leow and Crook (2014) developed an intensity model whose output was a matrix of transition probabilities between each pair of loan states within a period of six months. To achieve this result however, they had to make an adjustment to the intensity model to approximate the continuous time model to a discrete one since the observations they had were monthly observations and the dataset was not voluminous enough to justify the use of a continuous time model. The credit card default prediction challenge is more of a discrete-time problem since

observations are based on discrete time and customer response is also based on the same discrete times set for them (monthly deadlines).

2.4.4 Machine Learning Approaches

Machine learning applies computer algorithms and statistical techniques to build models that computers can use to solve problems without receiving explicit instructions (Rokad, 2019). In the analysis of credit card loan defaults, various approaches have been used in the past depending on the nature of the problem and intended outcomes. These approaches can generally be grouped into two categories:

i. Classification problems – simply classifying accounts as defaulters or non-defaulters after the time period for determination of default
ii. Prediction problems – using independent variables describing customers to forecast the customers' loan status after a certain duration. These also include studies which deal with intensity modeling of credit default, where a survival function is defined to calculate the probability of a customer's account "surviving" (remaining solvent) up to a certain time, *t.* These models are based on conditional probabilities of default (Lando, 2004).

Machine learning techniques are a good fit for modeling credit risk and can be used in various dimensions to obtain desired outcomes. Popular machine learning approaches used in analysis of credit risk in past studies are described below.

2.4.4.1 Decision Tree-Based Approaches

Islam (2018) applied the Extremely Randomized Trees (ET) algorithm on the Taiwan bank dataset, and developed a classification model that was based on a combination of metrics obtained from Online Analytics Processing (OLAP) data and Online Transaction Processing (OLTP) data. The latter involved combining the risk probability from archived data in a warehouse (OLAP) with the risk probability of a current transaction (OLTP). The idea of separating the data as OLAP and OLTP for analysis was needed because customer credit card data dates back many years, and such is normally warehoused so that transaction processing for new transactions may be efficient. The OLTP data is data that is as a result of transactions occurring within a selected small window of time in the dataset. The risk probability of OLAP data would first be precomputed and stored separately as a unit figure while that from OLTP would be computed in real-time as transactions occurred and then would be combined with the precomputed risk to arrive at the total risk factor. Figure 2.1 shows a flowchart of how the processing occurred.

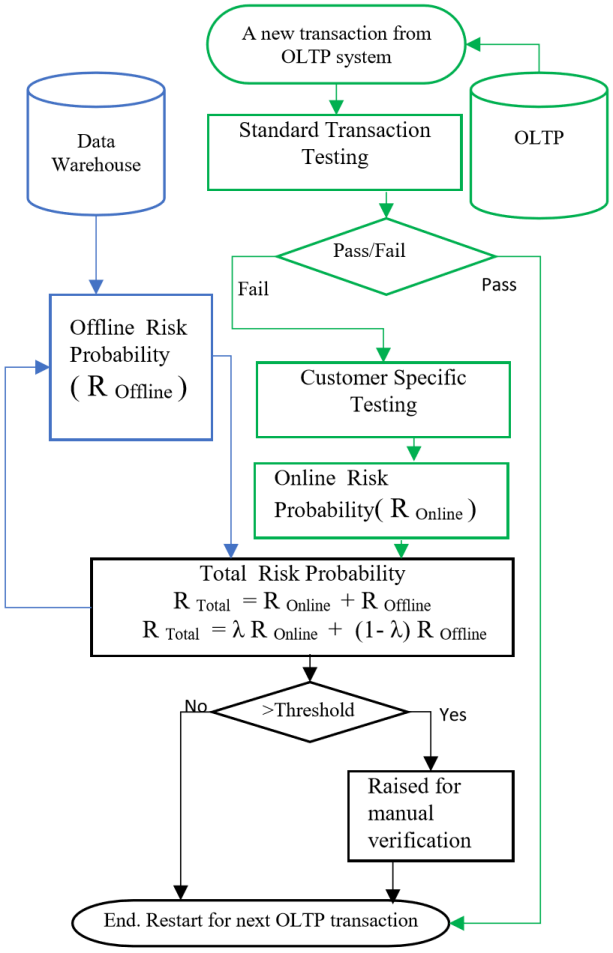

Figure 6.5: Flowchart of Risk Computation Process (Islam (2018))

A new transaction occurring in the OLTP system was first passed through a standard transaction testing process to check whether it deviated from any standard rules for performing transactions. If it passed, then no further testing would be done and the system would wait for the next transaction. If it failed, then the transaction would be subjected to customer specific testing which took into account customer specific rules to measure the risk probability.

Risk from offline data (OLAP) would be recalculated on a monthly basis to capture updates such as credit limit changes. Following this recalculation, the total risk would be updated only if the newly calculated total risk value was larger than what the system already had as the total risk. Note that the online risk probability was not, strictly speaking, calculated in real time but instead the algorithm was run on a batch of transactions whose number could be predetermined and varied. This is because calculating it in real-time (upon occurrence of a new transaction) would lead to a performance cost to transaction processing. The aim was to analyze the transactions as they occur and classify the account as a “good” account or a “bad” account

which would indicate whether the customer would make good on the value of that transaction by repaying their loan. The transactions under consideration were the purchase transactions only.

Under this approach, the formula for total risk was given as

$R_{Total} = \lambda R_{Online} + (1- \lambda)R_{Offline}$ where λ is the risk factor.

Risk probability from online data and offline data may carry different weights as assigned by the researcher. For example, one may give 40% weight to offline data and the remaining (60%) to online data. The figures to use would depend upon the company or organization which is relying upon this data, but Islam (2018) found that between 45% and 50% for online data and the remaining percentage for offline data would work best.

Islam (2018) also experimented with other classifiers, such as Naïve Bayes, J48, Rotation Forest and Extremely Randomized Trees on the offline data. Of all these, the findings were that Extremely Randomized Trees outperformed other algorithms in terms of accuracy, recall, precision, and F-score. The following table shows the analysis of results obtained for each algorithm type.

Table 2.1: Algorithm Accuracy, Precision, Recall and F-Score (Adapted from Islam (2018))

Algorithm	**Accuracy**	**Precision**	**Recall**	**F-score**
k-Nearest Neighbor	0.806	0.800697	0.806	0.793911
Support Vector Machine	0.776	0.76768	0.776	0.769471
Random Forest	0.936	0.936208	0.936	0.934927
Naïve Bayes	0.75	0.754993	0.75	0.75219
Gradient Boosting	0.865	0.864152	0.865	0.859669
Extremely Randomized Trees	**0.954**	**0.953856**	**0.954**	**0.953635**

2.4.4.2 Artificial Neural Networks

Kumar, Goel, Jain, Singhal and Goel (2018) analyzed credit risk using a multilayer perceptron (MLP) model of deep neural network on a dataset of 9578 instances to obtain a prediction accuracy of 93%. The model was implemented using a multilayer perceptron with 2 hidden layers of 20 nodes each, trained on 1000 epochs. The input layer of the model had 18 perceptrons which were activated using the function $y(v_1) = (1 + e^{-v_i})^{-1}$. This showed that neural networks can also be applied in credit risk modeling and with good results. The context

of application could be in classification of instances/accounts or prediction to forecast default events.

Artificial Neural Networks are mathematical simulations of the biological neural network, intended to imitate the working of the human memory (Bacham & Zhao, 2017). Neural networks comprise input and output layers, together with a hidden layer that transforms the inputs into a form that is usable by the output layer (Dormehl, 2019). They apply a technique called back propagation, which adjusts hidden layers of neurons until the output matches what the programmer intends. In this way, the ANN becomes optimal for that training dataset.

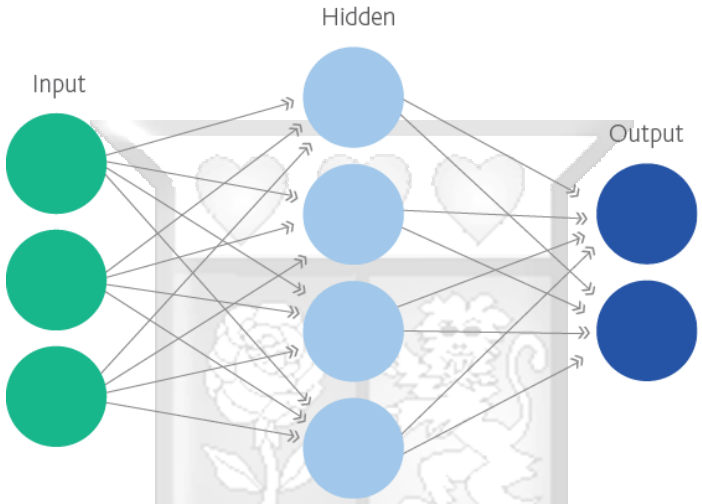

Figure 7.6: Artificial Neural Network (Adapted from (Bacham & Zhao, 2017))

2.4.4.3 Support Vector Machines (SVM)

Obare and Muraya (2011) performed a comparison between the logistic regression model and support vector machines for the modeling of credit risk for loan applicants of Equity Bank ltd, a bank in Kenya. The study tested both the linear SVM kernel and radial SVM kernel all on the same data and realized accuracy levels of 73% for logistic regression, 78% for the radial SVM and 86% for the linear SVM model. From this study the SVM models were a better fit for the data at hand and the problem defined as compared to the logistic regression model.

Support Vector Machines have proved successful in modeling of credit risk particularly due to their ability to support linearly non-separable scenarios whereas other methods such as discriminant analysis and logit analysis work well if the data is linearly separable (Chen, Hardle & Moro, 2011). The latter are however more interpretable in that it is possible to directly obtain the significance of variables in the study in determining the outcomes observed from the study.

A support vector machine (SVM) is a supervised learning model which, given training data, outputs an optimal hyperplane which categorizes new examples (Patel, 2017). In two dimensional space, the hyperplane is a line dividing a plane in two parts whereby each class lays in either side such that the plane has the maximum distance between the data points of both classes (Gandhi, 2018). Hyperplanes are decision boundaries that help to classify the data points, and data that falls on either side of the hyperplanes can be attributed to different classes, hence maximizing the distance between data points provides reinforcement so that future data points can be classified with more confidence (Gandhi, 2018).

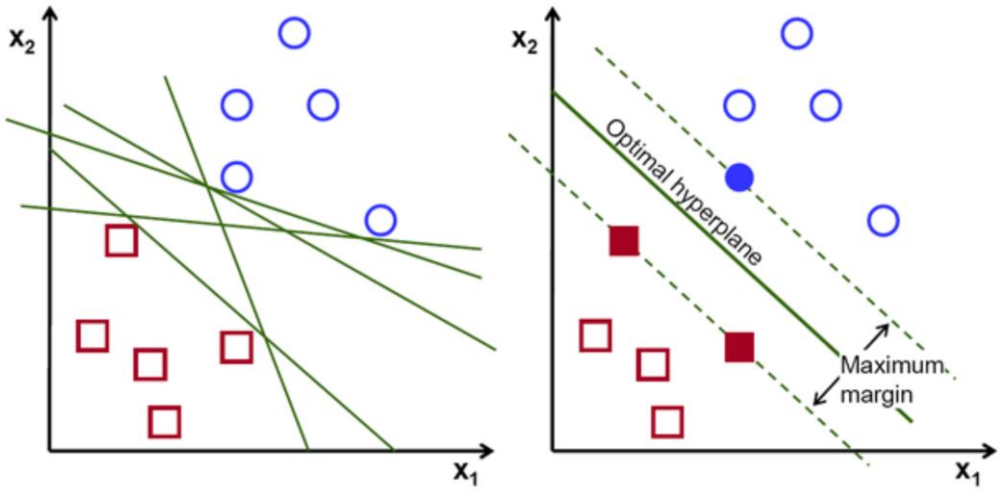

Figure 8.7: Possible SVM hyperplanes (Gandhi, 2018)

As Gandhi (2018) states, SVM finds a hyperplane in an N-dimensional space. If the number of input features is 2, then the hyperplane is a line. If the number of input features is 3, then the hyperplane is a two-dimensional plane.

2.5 Conceptual Framework

The developed model entailed use of credit card data that contained cardholders' personal characteristics and the transactions that they performed over a course of six months. The data was cleaned and important variables selected so that they can be used in the training of the logistic regression model. In particular, the required variables were those considered as a cardholder's personal characteristics and those which showed a customer's bill amounts and their repayments.

The cleaned data was then split into two, a training set and a test set, the training data being used to fit the model and the test data being used to validate the model. The model would then provide a prediction of a customer's loan status during the specified period. Figure 2.7 shows the conceptual framework.

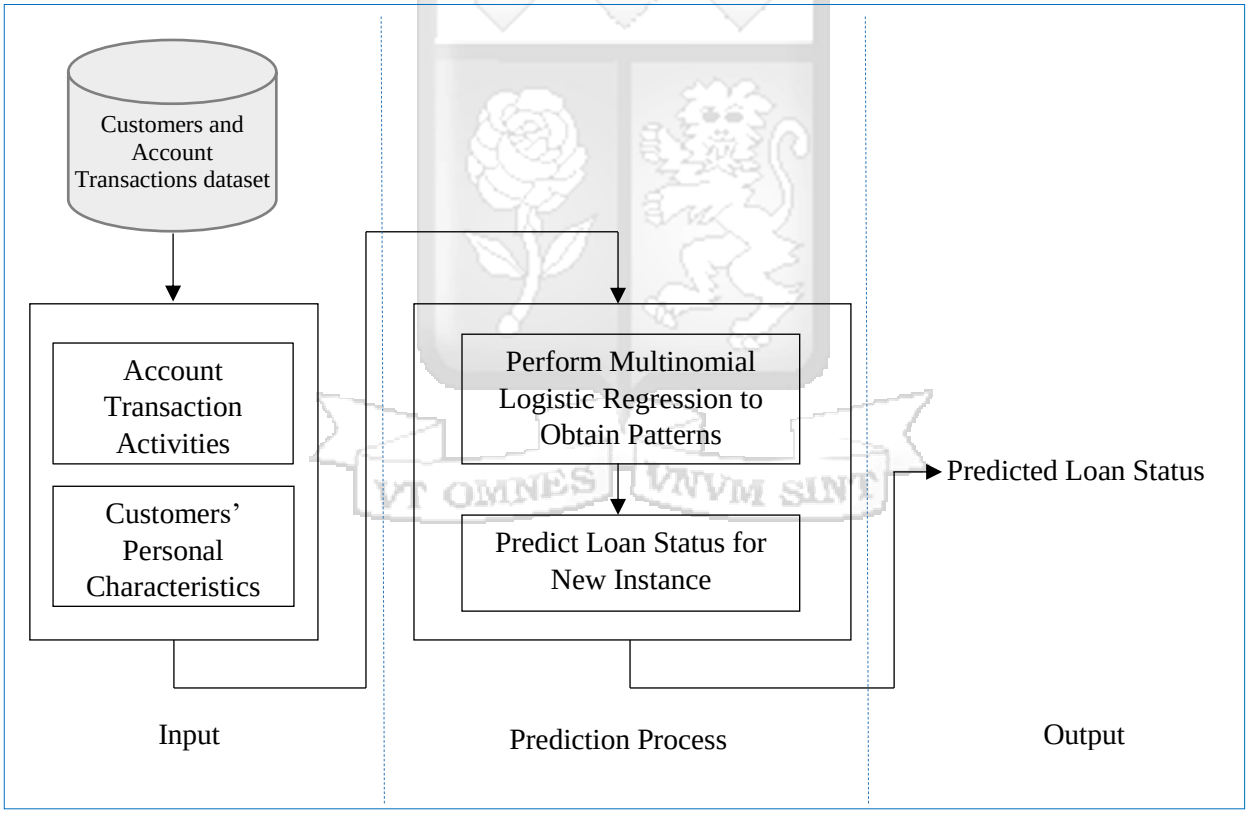

Figure 9.8: Conceptual Framework

Chapter 3: Research Methodology

3.1 Introduction

This study aims at developing a model that predicts the loan status of a credit cardholder given their sociodemographic (personal characteristics) and transactional data for a selected time period. The methodology required to achieve this outcome is hence detailed in this section.

3.2 Research Design

Quasi-experimental research design was used in this study to gain a cause-effect understanding of the relationship between variables under study and the outcome of credit card loans after specified time durations. Those variables would then be used in prediction of loan states in different customer contexts. This entailed determining which independent variables were most significant in predicting the loan status and how significant those variables were. Significance of the variables was assessed by establishing the impact that unit changes to the independent variables had to the dependent variable.

3.3 Model Development

The following steps were involved in development of the model

i. Obtaining the data
ii. Preprocessing the Data
iii. Developing the Model
iv. Validating the Model

3.3.1 Obtaining the Data

This research relied on the Taiwan Bank dataset which has been made publicly available for research purposes and which contains cardholders' account information. This includes their transactions for a period of six months and their loan statuses during those periods. Obtaining such data from financial institutions is difficult due to the sensitive nature of the data contained in them, and the privacy implications that are present for the cardholders.

Required data for this study was data related to cardholders' sociodemographic attributes such as age, marriage status and education level and also their transaction data, that is, the transactions that they performed that were related to their credit cards. With this data, it was possible to build a model for prediction of default of credit card loans.

3.3.2 Preprocessing the Data

This entailed listing and describing the available variables in the data, establishing the data types of values within the instances of that data and observing conformance to implied data type requirements for variable values in all instances that were selected to be part of the study.

Data exploration was done through calculation of averages and counts for relevant variables, maximum values and minimums, medians and modes in order to get better sense of the data and the variables therein, and also to observe the balance/imbalance portrayed by the data with respect to the target variable under consideration.

Values contained in the different fields in the data were also checked to see if there were any apparent patterns or notable features, for example, variables which appeared to be categorical variables (and those which were by their nature expected to be categorical variables) were converted into factors for better analysis. For factor variables which were nominal in nature, indicator/dummy variables were created which contained binary values indicating the presence or absence of a categorical effect. Categorical variables which were ordinal in nature were retained as numeric types.

Data preprocessing also included handling of null values contained in the data and generation of the final dataset that would be used for the study.

3.3.3 Developing the Model

The model was developed using the nnet package which is used for developing multinomial log-linear models. The model parameters were set, and settings for the outcome variable configured (the number of levels of the independent variable) and the model trained on the training data.

3.3.7 Validating the Model

The developed model was validated through 10-fold cross validation to determine the accuracy of the model in predicting the loan status for different customers and for specified time periods.

3.4 System Development Methodology

This research utilized an agile development methodology, which is both iterative and incremental, in developing a system which implements the default prediction model. Agile methodology allows for developing of initial system designs and iteratively adding onto the designs as system features are refined (Gonçalves, 2019). This is suitable for a machine

learning project as the algorithms are fine-tuned based on the data and problem domain. Multiple iterations hence needed to be factored into the selected methodology.

The Agile process breaks down a larger software project into smaller sub-projects/modules which can be implemented in iteration (Gonçalves, 2019). At the onset, functional system modules were identified and the design and implementation efforts channeled towards developing these subsystems in increments and iterations. Prioritization was based on covering the base functionality and then appending helper functions in the later iterations. In this way, the final software product was developed and delivered in a faster way.

3.5 System Analysis

A structured systems analysis and design approach was adopted in performing requirements analysis and developing the designs upon which the system would be implemented. System requirements for the system were based on providing a solution in the credit card loan default prediction problem domain and enabling target users to use the system in a simple and efficient manner. This system was intended to be used in banking and other financial institutions and the main users of the system are the loan officers and collectors.

System analysis entailed developing of the user requirements and also translating the requirements to a logical specification of what the system was supposed to do. This was important in forming the foundation upon which modular design would be achieved and hence the incremental and iterative builds which led to the delivery of the final working solution. The output of the system analysis stage included:

i. A requirements catalogue
ii. A data catalogue
iii. The process specification, which included the logical data model

3.7 System Design

Details of the implementation at a logical level were specified during system design to form the basis upon which the system modules would be constructed. Designs were specified at a high level to avoid excessive detailing which would have been restrictive during the actual implementation and which would have led to loss of time in scenarios where changes to designs were required. The design process itself was also incremental, since it was approached in a modular fashion, with core system functionality being considered first and additional sections designed later and added to the existing designs. This therefore required that initial designs be high level and also open for extension. The output of the system design processes included:

i. A logical data model
ii. Logical process model
iii. A database schema

3.8 Data Collection

The study utilized secondary data which is quantitative in nature. The data was historic data of credit card customers of Taiwan Bank covering the period from April 2005 to September 2005 and which contained customer profiles (their attributes and their transactions for that period). Secondary data was found to be suitable due the difficulty in obtaining first hand data directly from financial institutions (sharing of credit card data could have security and privacy concerns) and also to hasten the acquisition of data for the study. The Taiwan dataset has been made public to the research community and is also relatively more recent in comparison to other dataset that can be used for credit risk default modeling. In addition, it contained majority of the attributes that were required for the study, and in particular the customers' personal characteristics and their transactions history. This made it a suitable dataset for use in this study.

The dataset contains 30,000 instances, where it is not possible to identify the actual clients implicated in the data records. Attributes that would otherwise identify the clients were excluded, however, the rest of the attributes were present and well described. The dataset source was UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science.

3.9 Data Analysis

In order to achieve the objective of creating a prediction model for loan defaulting, an analysis on the cardholder characteristics and account transaction activities data had to be performed so as to reveal the patterns exhibited in the data. This was done through regression analysis to determine how each attribute would contribute to the outcome, that is, the loan status at the end of the observation period.

Initially, descriptive analysis was performed to reveal the data distribution and associations amongst variables. Box-plots were plotted for quantitative variables such as bill amounts to show the quartiles distribution and as a basis for identification of outliers. Bar plots were used for categorical variables, SEX, EDUCATION and MARRIAGE to show their distributions and possibility of non-documented values.

Logistic regression analysis was then performed to generate the regression line that summarizes the relationship between predictor variables and the odds of defaulting (transformed to log of

odds of defaulting). This could then be used to predict the probability of defaulting of a loan give the values of the predictor variables. This analysis included determining the coefficient values of the covariates, which also showed the effect of unit changes of the predictor values in determining the outcome of the loan relative to other predictors.

3.10 Research Quality

3.10.1 Reliability

Reliability of results obtained was estimated using the measures of model accuracy, recall and precision. 10-fold cross validation was used to determine the mean accuracy, the mean recall and the mean precision (for all folds in cross validation).

The accuracy of the model used was obtained using the following formula:

$$\text{Accuracy of model} = \frac{\text{Correctly predicted observations}}{\text{Total number of observations}} * 100$$

Equation 3.1: Formula for Computing Accuracy of classifier model (Adapted from (Browlee, 2014))

The precision metric was used to show the exactness of the model. It shows the proportion (percentage) of the model's results which can be relied upon (Saxena, 2018). To obtain the model precision for a multinomial distribution, the precision is first obtained with respect to each class in the multiclass target variable, then the average of the precisions is obtained by dividing the sum of individual precisions (per class) by the number of classes. The following formula was used for obtaining the precision of the model:

$$\text{Precision} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}} * 100$$

$$\text{Precision of Multinomial model} = \frac{\sum \text{Precision for each class}}{\text{Number of classes}} * 100$$

Equation 3.2: Formula for Computing Precision of Model (Adapted from (Browlee, 2014))

The model's recall value is a measure of completeness of the model, which shows how much of the relevant observations have been correctly predicted by the model. To obtain the model recall for a multinomial distribution, the recall is first obtained with respect to each class in the multiclass target variable, then the average of the recalls is obtained by dividing the sum of

individual recalls (per class) by the number of classes (Saxena, 2018). This was calculated using the following formula:

$$\text{Recall} = \frac{\text{Number of positive predictions}}{\text{Total number of positive class values}} * 100$$

$$\text{Recall of Multinomial model} = \frac{\sum \text{Recall for each class}}{\text{Number of classes}} * 100$$

Equation 3.3: Formula for computing recall of model (Adapted from (Browlee, 2014))

3.10.2 Validity

The research endeavored to use and present training data as it was, allowing for data pre-processing so that the model could work as intended, but without other manipulation in such a way as to yield biased results and all results were presented and reported as they were without unwarranted modification.

3.11 Ethical Considerations

The research proposal for this study was reviewed and approved by the Strathmore University Institutional Ethics Review Committee (SU-IERC). The approval for conducting the study was issued on 14th January 2020. The proposal was also reviewed by the National Commission for Science, Technology and Innovation (NACOSTI) and an approval granted for the study on 10th March 2020.

Chapter 4: System Design and Architecture

4.1 Introduction

This study focuses on developing a prediction model for default of credit card loans. A consideration of the parameters required for the development of the model and of the system which implements that model will be included in this chapter and details of how those parameters will be translated to the desired final outcomes will be highlighted. The design and architecture of the system which will implement the default prediction model according to the conceptual framework shown in section 2.5 will be explained in detail including the software entities that make up the system, the data models, processes and interactions involved in the prediction of loan defaulting.

4.2 Data Analysis

4.2.1 Dataset Source

The study utilized secondary data sourced from UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science. The data was historic data of credit card customers of Taiwan Bank covering the period from April 2005 to September 2005 and which contained customer profiles (their attributes and their transactions for that period). Secondary data was found to be suitable due the difficulty in obtaining first hand data directly from financial institutions (sharing of credit card data could have security and privacy concerns) and also to hasten the acquisition of data for the study. The Taiwan dataset has been made public to the research community and is also relatively more recent in comparison to other dataset that can be used for credit risk default modeling. In addition, it contained majority of the attributes that were required for the study, and in particular the customers' personal characteristics and their transactions history. This made it a suitable dataset for use in this study.

The dataset has also been used in other studies such as Islam (2018) and Drugov (n.d.) which managed to successfully develop classification models for defaulters and non-defaulters of credit card loans, showing that the data is reliable for use in such studies.

4.2.2 Description of the Dataset

The Taiwanese dataset contains 25 variables as described by (Ma, 2020). The following table shows the variables and their descriptions:

Table 4.1: Taiwanese data description (Adapted from Ma, 2020))

	Variable	**Description**
1	ID	The client ID
2.	LIMIT_BAL	Amount of given credit in NT Dollars
3.	SEX	Gender (1 = M, 2 = F)
4.	EDUCATION	1 = Graduate school, 2 = University, 3 = High school, 0, 4, 5, 6 = Others
5.	MARRIAGE	Marital Status (1 = Married, 2 = Single, 3 = Divorce, 0 = Others)
6.	AGE	Age in years
7.	PAY_0	Repayment status in September, 2005 -2 = No consumption -1 = Paid in full, 0 = The use of revolving credit, 1 = Payment delay for one month, 2 = Payment delay for two months, … 8 = Payment delay for eight months, 9 = Payment delay for nine months and above
8.	PAY_2	Repayment status in August, 2005 (same scale as in PAY_0)
9.	PAY_3	Repayment status in July, 2005 (same scale as in PAY_0)
10.	PAY_4	Repayment status in June, 2005 (same scale as in PAY_0)
11.	PAY_5	Repayment status in May, 2005 (same scale as in PAY_0)
12.	PAY_6	Repayment status in April, 2005 (same scale as in PAY_0)
13.	BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
14.	BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
15.	BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
16.	BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
17.	BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
18.	BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
19.	PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)
20.	PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
21.	PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)

22.	PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
23.	PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
24.	PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
25.	default.payment.next.month	Client's behavior (1 = default, 0 = no default)

4.2.3 Data Cleaning

This process involved identification of instances with incorrect entries, such as non-conformant or undescribed values within categorical variables, null/missing values within the data and duplicates in the data. The rule established for handling such entries was that records with undescribed values for categorical variables and those with missing values in their fields would be removed in entirety. For duplicate entries, only the repeated entry would be removed from the dataset.

An analysis of the data revealed that all variables with categorical data conformed to the described levels in that there were no values that were not provided in the data description adapted from (Ma, 2020). Also, there were no missing values in the entire dataset and no duplicate records in the dataset. The check for duplicate records was done by comparing combined values/fields for entire records, not just single variables since single variables may rightly have repeated values.

4.2.4 Descriptive Analysis

A structural analysis of the dataset gives a look at the makeup of the dataset based on its variables. All variables are either numeric or integers with the integer variables being categorical or auto-increment variables (ID) and the numeric variables being amounts.

```
'data.frame':   30000 obs. of  25 variables:
 $ ID                        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ LIMIT_BAL                 : num  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000 ...
 $ SEX                       : int  2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION                 : int  2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE                  : int  1 2 2 1 1 2 2 2 1 2 ...
 $ AGE                       : int  24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_0                     : int  2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2                     : int  2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3                     : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4                     : int  -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5                     : int  -2 0 0 0 0 0 0 0 0 -1 ...
 $ PAY_6                     : int  -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1                 : num  3913 2682 29239 46990 8617 ...
 $ BILL_AMT2                 : num  3102 1725 14027 48233 5670 ...
 $ BILL_AMT3                 : num  689 2682 13559 49291 35835 ...
 $ BILL_AMT4                 : num  0 3272 14331 28314 20940 ...
 $ BILL_AMT5                 : num  0 3455 14948 28959 19146 ...
 $ BILL_AMT6                 : num  0 3261 15549 29547 19131 ...
 $ PAY_AMT1                  : num  0 0 1518 2000 2000 ...
 $ PAY_AMT2                  : num  689 1000 1500 2019 36681 ...
 $ PAY_AMT3                  : num  0 1000 1000 1200 10000 657 38000 0 432 0 ...
 $ PAY_AMT4                  : num  0 1000 1000 1100 9000 ...
 $ PAY_AMT5                  : num  0 0 1000 1069 689 ...
 $ PAY_AMT6                  : num  0 2000 5000 1000 679 ...
 $ default.payment.next.month: int  1 1 0 0 0 0 0 0 0 0 ...
```

Interesting to note is the LIMIT_BAL variable whose levels begin from 10,000 and go up to 1,000,000, advancing upwards by 10,000 NT dollars each time except for one instance where the LIMIT_BAL value was 16,000. This variable reveals the level of confidence that the bank had initially placed on the customer's ability to repay their loan. The dataset does not have information about the changes in limit balance over time due to the customer's repayment behavior, hence that aspect of the study will be left out.

A statistical summary of the independent variables used in the dataset is shown in Table 4.2.

Table 4.2: Statistical Summary of Independent Variables in the Dataset

Variable	Mean	Median	Min	Max
Limit_Bal	167,484.3	140,000	10,000	1,000,000
Sex	-	-	-	-
Education	-	-	-	-
Marriage	-	-	-	-
Age	35.49	34	21	79
Pay_0	−0.0167	0	-2	8
Pay_2	−0.1338	0	-2	8
Pay_3	−0.1662	0	-2	8
Pay_4	−0.2207	0	-2	8
Pay_5	−0.2662	0	-2	8
Pay_6	−0.2911	0	-2	8
Bill_Amnt1	51,223.33	22,381.5	-165,580	964,511
Bill_Amnt2	49,179.08	21,200	-69,777	983,931
Bill_Amnt3	47,013.15	20,088.5	-157,264	1,664,089
Bill_Amnt4	43,262.95	19,052	-170,000	891,586
Bill_Amnt5	40,311.4	18,104.5	-81,334	927,171
Bill_Amnt6	38,871.76	17,071	-339,603	961,664
Pay_Amnt1	5,663.58	2,100	0	873,552
Pay_Amnt2	5,921.16	2,009	0	1,684,259
Pay_Amnt3	5,225.68	1,800	0	896,040

Pay_Amnt4	4,826.08	1,500	0	621,000
Pay_Amnt5	4,799.39	1,500	0	426,529
Pay_Amnt6	5,215.50	1,500	0	528,666

Summaries for Sex, Education and Marriage were not obtained since these variables are nominal categorical variables, hence the summaries would be of no values. Summaries for Pay_0 to Pay_6 were however obtained since these were considered ordinal variables, where graduating from the lowest to the highest value implies graduating from lower default severity to higher default severity. It was therefore important to view the summaries for severity levels associated with this variable.

The following histograms show the count distributions of Sex, Education and Marriage variables. From the analysis, it was found that no undescribed values were present in the categorical variables and that all levels of the variables had a representation in the dataset.

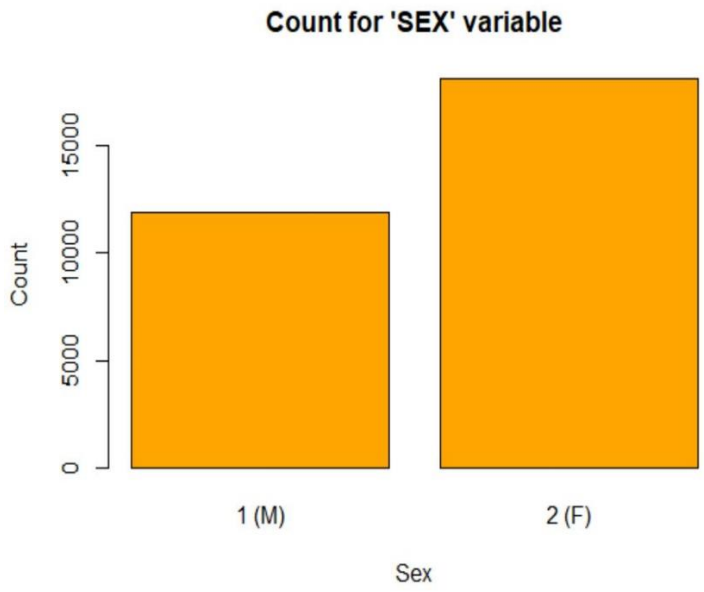

Figure 4.1: Count Histogram for ‘SEX’ Variable

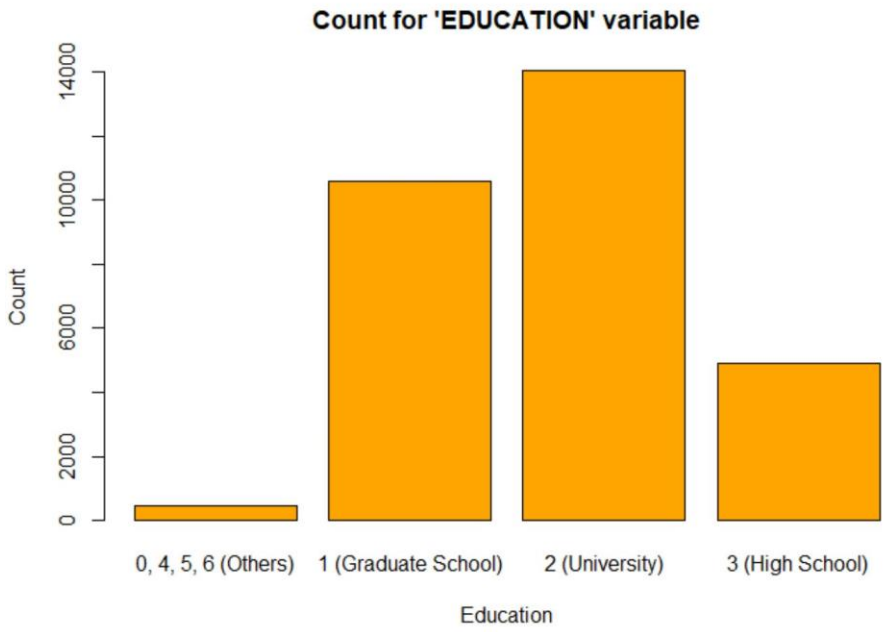

Figure 4.2: Count Histogram for ‘EDUCATION Variable

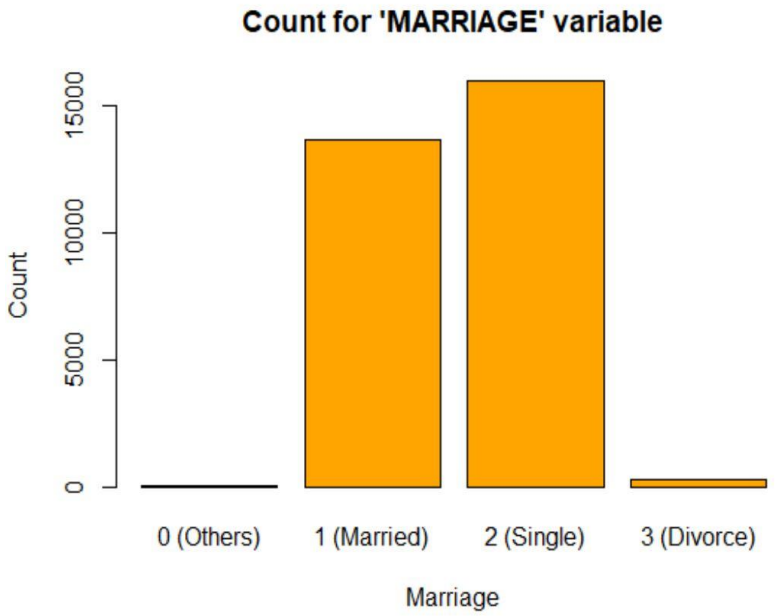

Figure 4.3: Count Histogram for ‘MARRIAGE’ Variable

Detection of outlier values was done through the use of boxplots. Figures 4.4 and 4.5 show the boxplots for Bill Amount and Pay Amount variables and are each followed by a description of how outliers were determined for the variables.

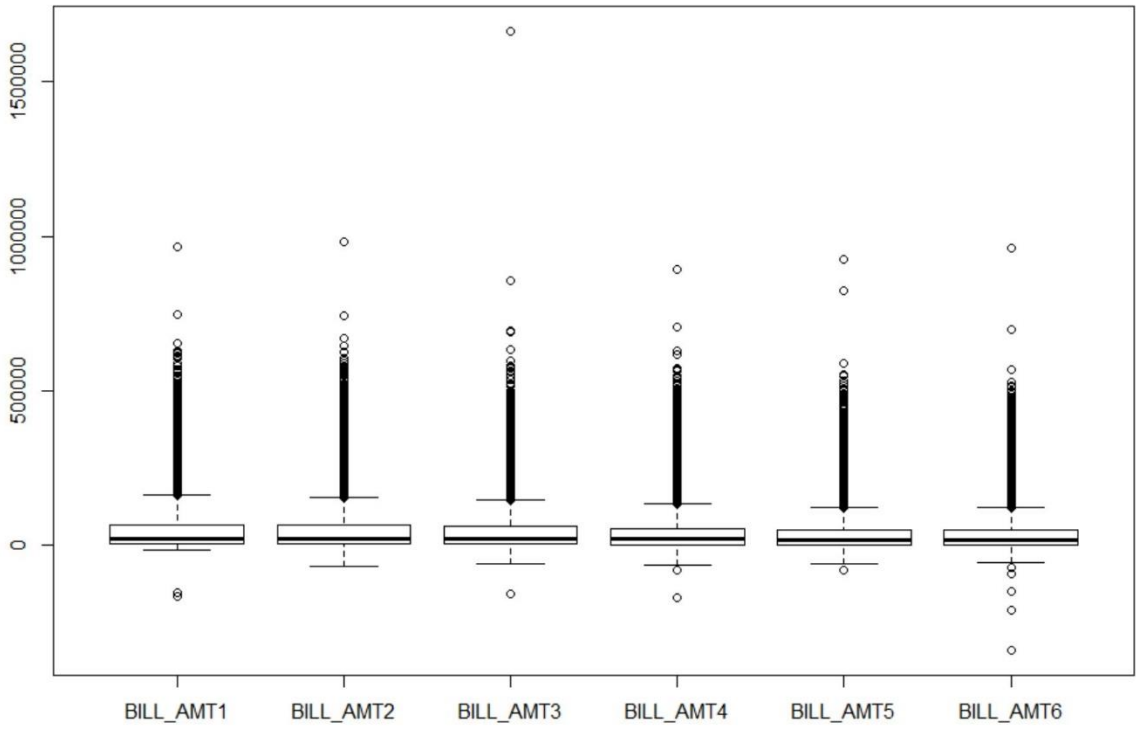

Figure 4.4: Box-plot for 'BILL_AMNT' Variables

The BILL_AMNT variables were depicted the amount of bill statement during the respective months. Since these can actually be considered as repeated observations of the same variable (per cardholder), then for the purpose of outlier identification, they were analyzed using the same boxplot. In addition, and to minimize the loss of information due to elimination of outliers, a cut-off value was set for these variables based on the general trend observed in the six variables. Consequently, it was assumed that for all 'BILL_AMNT' variables, any amount exceeding 1,000,000 and any amount going below *min(BILL_AMNT4)*, which was the value -170,000, was considered to be an outlier.

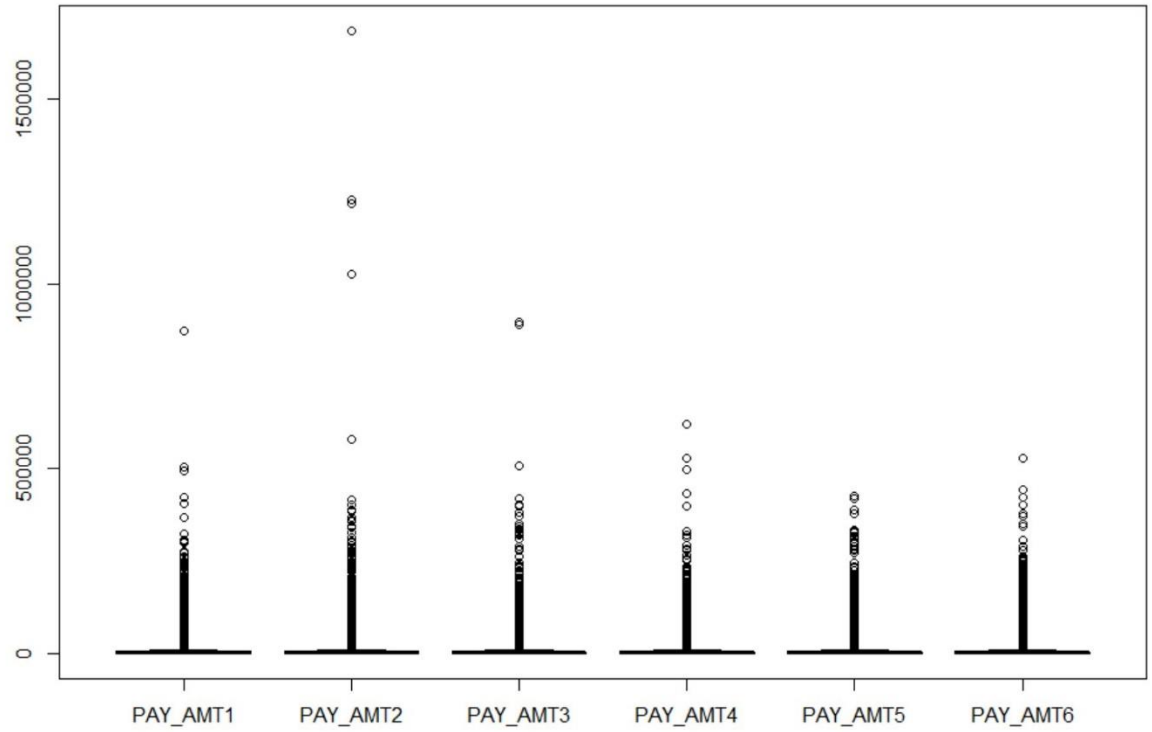

Figure 4.5: Box-plot for 'PAY_AMNT' Variables

For the 'PAY_AMNT' variables, the same approach was takes for outlier detection as the one taken for 'BILL_AMNT' variables. The general trend of the six variables set the upper cut-off point at *max(PAY_AMNT4)* which was the value 621,000. Any observation exceeding this amount was considered an outlier. On the lower side, no observation had a value lower than 0, hence there were no outliers based on this (taking 0 as the minimum).

After the exclusion of outlier values was done, a sample size of 29,993 remained (7 observations were dropped), which was the data used for training and validating the model.

4.3 Requirements Analysis

In order to fully serve its purpose, the developed system sought to address two categories of requirements: functional and non-functional requirements. Functional requirements entail performance of required tasks by the system. They refer to what the system is supposed to do to fulfil the needs of the user (Eriksson, 2012). Non-functional requirements are the quality attributes that refer to how well a system is performing its objectives (Spacey, 2017).

4.3.1 Functional Requirements

i. The system should allow loan officers to upload credit card data (cardholders' personal attributes and their transaction activities) in form of .csv file

ii. The system should determine how cardholders' personal attributes and their transaction activities determine the rate of defaulting for credit card loans
iii. The system should provide the loan officers with the prediction of loan statuses for customers given their personal characteristics and their transaction activities

4.3.2 Non-functional Requirements

The system operates using sensitive customer information which could have significant concerns to the business is handled improperly. Also, the system was intended to be used by loan officers who may have intermediate expertise in the use and understanding of computer systems at the least. In order to achieve these requirements, the system was therefore expected to have:

i. High data privacy – this would ensure that the data would be accessed by only the intended users and that sharing of system data with third parties would be done in an authorised way and with proper trails. The system should not allow data to be disclosed or availed to unintended users.
ii. High data security – the system should guarantee high confidentiality, availability and integrity of data.
iii. High reliability – for the system to serve its purpose, high availability and correctness of results has to be guaranteed to its users
iv. High performance – the system has to have low response time in giving of outputs and has to be economical in its usage of resources

4.4 System Architecture

The system architecture is as shown in Figure 4.6. It constitutes a prediction engine which will run a prediction algorithm optimized through the training data and validated using the test data. New/unlabelled instances will then be assigned a prediction of the loan status for the requested period. The output is the status of the loan, that is, whether the cardholder will pay the amount in full or whether they will pay less than the minimum required repayment or more than the minimum, or whether the customer will not pay at all.

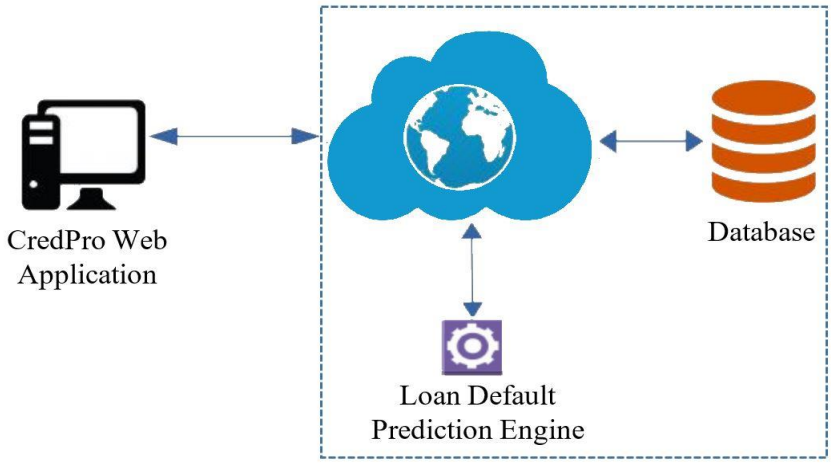

Figure 4.6: System Architecture

4.5 Context Diagram

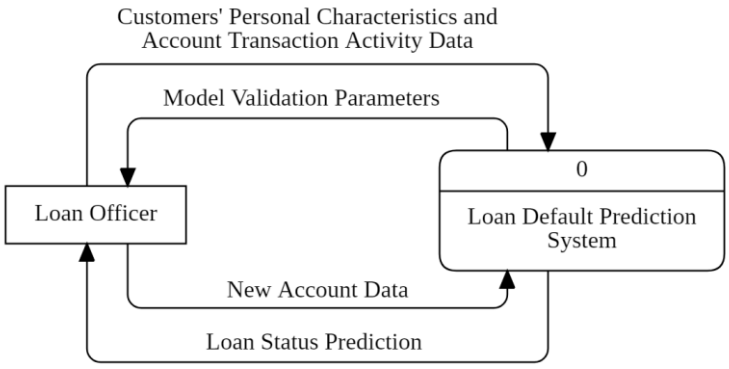

Figure 4.7: Context Diagram

The main external entities (system actors) are the loan officers. The loan officer is responsible for preprocessing the data to be uploaded to the system and training the model based on the supplied data. Preprocessing includes all preparations done to the data to make it ready for use in model building. This includes data cleaning and performing conformance checks of the system requirements for the data. Loan officers then input new account data and obtain predictions of their loan statuses during the specified period. They may also obtain predictions for existing accounts but for unobserved time periods.

4.6 Level 1 Data Flow Diagram

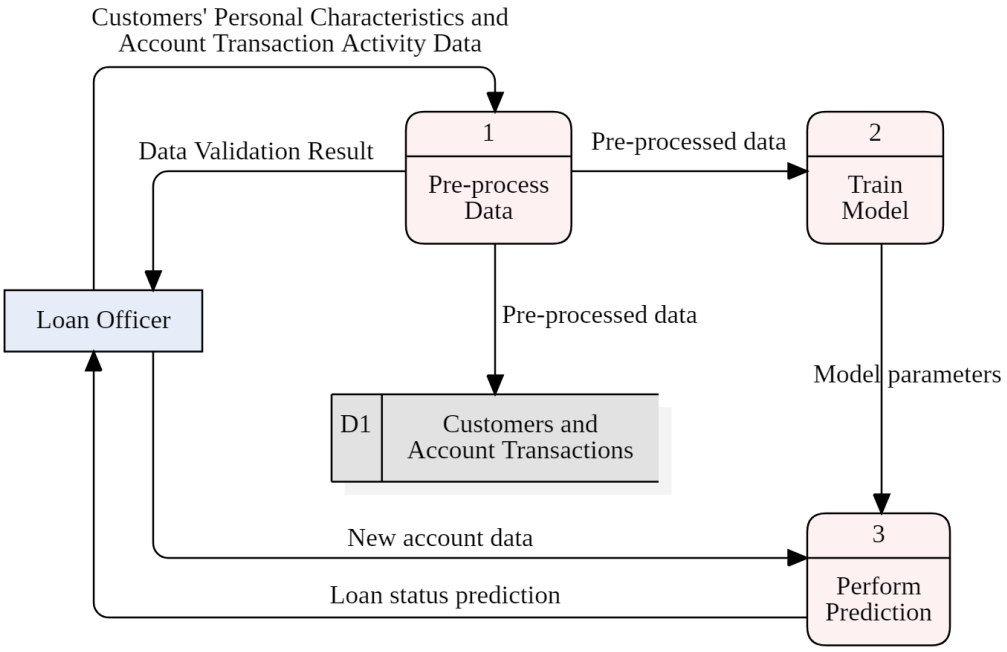

Figure 4.8: Level 1 DFD

4.7 Conceptual Data Model

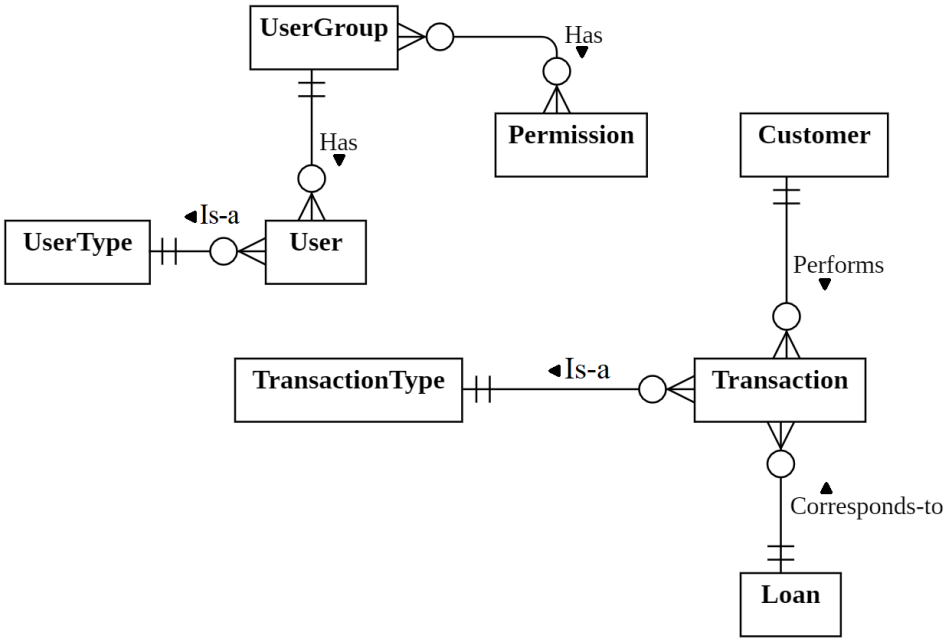

Figure 4.9: Entity Relationship Diagram

4.8 Database Schema

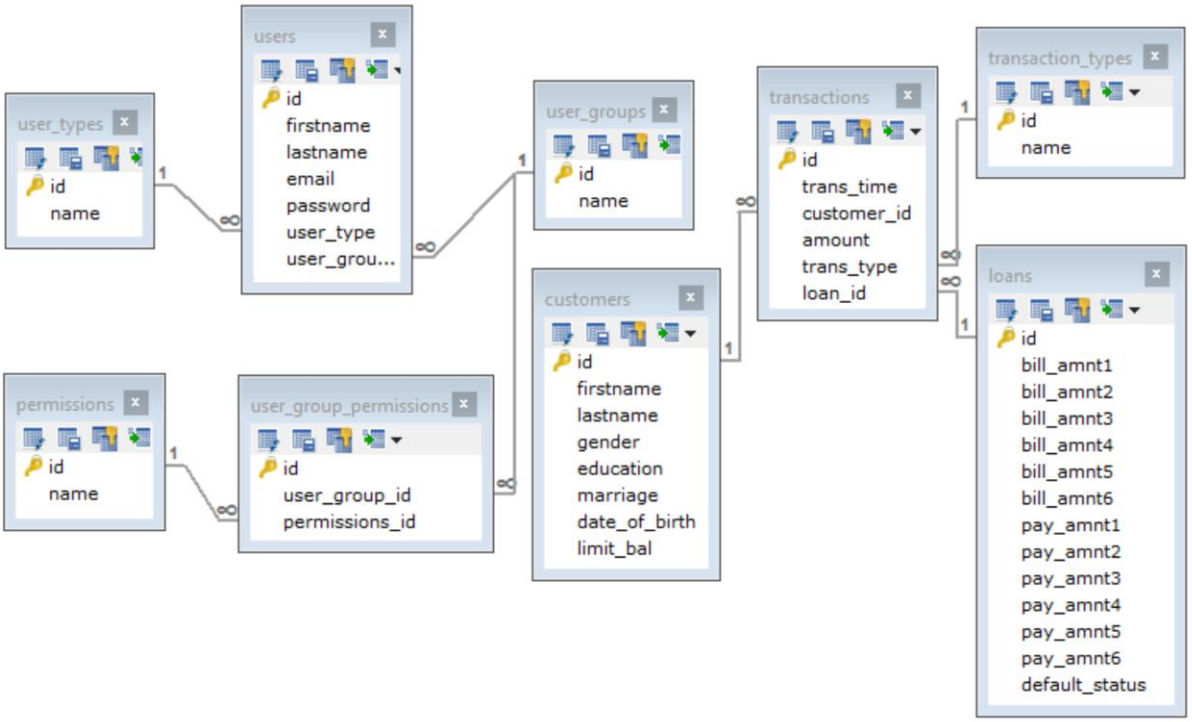

Figure 4.10: Database Schema

The diagram above shows the entity relationship model for the loan default prediction system. The schema covers the part of the system engaged in model building and prediction of loan status for cardholder accounts included in the loans and customers table. User management includes distinguishing the different types of system users (System Administrators and Loan Officers) and assigning permissions for system usage so that the users can only do what they are authorized to do. User authorization for system usage is provided for through email and password authentication (for end users of the system).

Chapter 5: Implementation and Testing

5.1 Introduction

The system comprised a high-level module that had an interface to the model which was a subsystem that implemented the prediction algorithm. The high-level module of the system implemented the user management functionality, data upload and preliminary validation of data, interface to the data storage engine, interface to the default prediction model and also provided the graphical user interface. This was a web application created using the Java Spring Boot Framework. The default prediction model was embedded to the system as a dynamically linked library, whose implementation was done in R language. The data used in the system was stored in a MYSQL database.

The web application enabled the users to log in to their accounts, upload training data to the system and input new instances of cardholder data to obtain predictions. For this to occur a call would be made to the R library which contained the default prediction model. System Administrators are the actors responsible for initial system configuration while loan officers would train the model and obtain predictions for new cardholder instances. The system would also give the total probability of a certain requested loan status for a certain month, a metric which can be used to depict the level of a bank's net risk portfolio with regards to credit card loans.

5.2 Model Development

The system relied on a multinomial logistic regression model to determine the factors which were relevant in determining loan outcomes and consequently to perform predictions of loan statuses given the required input values. The model was created using 'nnet' library of the R programming language. The following section shows the step-by-step approach taken in implementing the model.

5.2.1 Data Preprocessing

The required input was a dataset in CSV format. Before the data could be used in developing the model, it had to undergo a pre-processing phase which had the following steps:

i. Data Cleaning

   This entailed removal of statistically insignificant variables (such as the ID variable), removal of instances with missing values and removal of instances with undescribed values for the categorical variables. This was done as described in Section 4.2.3. Removal of outliers was done according to the description in Section 4.2.4.

The following R syntax was used to perform data cleaning:

```
#Remove 'repayment state variables
data <- subset(data, select = -c(PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6) )

#Remove the ID Variable
data <- subset(data, select = -c(ID) )

#Remove default.payment.next.month
data <- subset(data, select = -c(default.payment.next.month) )

#Remove outliers (BILL_AMNT values above 1,000,000)
data <- subset(data,BILL_AMT1 <= 1000000 & BILL_AMT2 <= 1000000 &
                                  BILL_AMT3 <= 1000000 & BILL_AMT4 <= 1000000 &
                                  BILL_AMT5 <= 1000000 & BILL_AMT6 <= 1000000 )

#Remove outliers (PAY_AMNT values above 621,000)
data <- subset(data, PAY_AMT1 <= 621000 & PAY_AMT2 <= 621000 &
                                  PAY_AMT3 <= 621000 & PAY_AMT4 <= 621000 &
                                  PAY_AMT5 <= 621000 & PAY_AMT6 <= 621000 )
#remove outliers (BILL_AMNT values below -170,000)
data <- subset(data, BILL_AMT1 >= -170000 & BILL_AMT2 >= -170000 &
                                  BILL_AMT3 >= -170000 & BILL_AMT4 >= -170000 &
                                  BILL_AMT5 >= -170000 & BILL_AMT6 >= -170000 )
```

Figure 5.1: Data Cleaning Algorithm

ii. Reshaping data to long (longitudinal) format

The original dataset was presented in a 'wide' format, where every instance contained six months' observations for a single cardholder. Before further analysis, this data had to be restructured in longitudinal format. This included restructuring the data in a stacked format so that variables 'BILL_AMNT1' to 'BILL_AMNT6' were collapsed into one variable named 'BILL_AMNT' and 'PAY_AMNT1' to 'PAY_AMNT6' were collapsed into one variable named 'PAY_AMNT' making it a longitudinal dataset where each record represented an individual customer's data for one month.

Figures 5.2 and 5.3 show the R syntax that was used to perform conversion to longitudinal format:

```r
get_repayment_rate <- function(bill, payment) {
  if(bill < 0) {
    return (2)
  }
  if(0 == bill) {
    #No consumption
    return(0)
  }
  if(payment < bill) {
    return (1)
  }
  return (2)
}
add_month_data <- function(month, record, prev_bill, prev_payment) {
  bill <- NA
  payment <- NA
  if( 4 == month ) { #April
    bill <- record$BILL_AMT6
    payment <- record$PAY_AMT5
  } else if( 5 == month ) { #May
    bill <- record$BILL_AMT5
    payment <- record$PAY_AMT4
  } else if( 6 == month ) { #June
    bill <- record$BILL_AMT4
    payment <- record$PAY_AMT3
  } else if( 7 == month ) { #July
    bill <- record$BILL_AMT3
    payment <- record$PAY_AMT2
  } else if( 8 == month ) { #August
    bill <- record$BILL_AMT2
    payment <- record$PAY_AMT1
  } else if( 9 == month ) { #Sep
    bill <- record$BILL_AMT1
    payment <- 0
  }

  dat <- data.frame(matrix(nrow = 0, ncol = 10))
  colnames(dat) <- c("LIMIT_BAL", "SEX", "EDUCATION", "MARRIAGE", "AGE", "PREV_BILL", "PREV_PAYMENT", "BILL", "PAYMENT", "TARGET")

  dat[1, "LIMIT_BAL"] <- record$LIMIT_BAL
  dat[1, "SEX"] <- record$SEX
  dat[1, "EDUCATION"] <- record$EDUCATION
  dat[1, "MARRIAGE"] <- record$MARRIAGE
  dat[1, "AGE"] <- record$AGE
  dat[1, "PREV_BILL"] <- prev_bill
  dat[1, "PREV_PAYMENT"] <- prev_payment
  dat[1, "BILL"] <- bill
  dat[1, "PAYMENT"] <- payment
  dat[1, "TARGET"] <- get_repayment_rate(bill, payment)

  return (dat)
}
```

Figure 5.2: Utility Functions for Conversion to Longitudinal Format

```r
data1 <- data.frame(matrix(nrow = 0, ncol = 10))
colnames(data1) <- c("LIMIT_BAL", "SEX", "EDUCATION", "MARRIAGE", "AGE", "PREV_BILL", "PREV_PAYMENT", "BILL", "PAYMENT", "TARGET")

for (i in 1:nrow(data)) {
  record = data[i,]
  #May Data
  data1 = rbind(data1, add_month_data(5, record, data[i, "BILL_AMT6"], data[i, "PAY_AMT5"]))
  #June Data
  data1 = rbind(data1, add_month_data(6, record, data[i, "BILL_AMT5"], data[i, "PAY_AMT4"]))
  #July Data
  data1 = rbind(data1, add_month_data(7, record, data[i, "BILL_AMT4"], data[i, "PAY_AMT3"]))
  #Aug Data
  data1 = rbind(data1, add_month_data(8, record, data[i, "BILL_AMT3"], data[i, "PAY_AMT2"]))
}
```

Figure 5.3: Algorithm for Conversion to Longitudinal Format

Tables 5.1 and 5.2 demonstrate the effect of restructuring the dataset in the long format.

Table 5.2: Customer and Transactions Dataset (Before Transformation)

ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
1	20000	2	2	1	24	3913	3102	689	0	0	0	0	689	0	0	0	0
2	120000	2	2	2	26	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000
3	90000	2	2	2	34	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000
4	50000	2	2	1	37	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000
5	50000	1	2	1	57	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679
6	50000	1	1	2	37	64400	57069	57608	19394	19619	20024	2500	1815	657	1000	1000	800
7	500000	1	1	2	29	367965	412023	445007	542653	483003	473944	55000	40000	38000	20239	13750	13770
8	100000	2	2	2	23	11876	380	601	221	-159	567	380	601	0	581	1687	1542
9	140000	2	3	1	28	11285	14096	12108	12211	11793	3719	3329	0	432	1000	1000	1000
10	20000	1	3	2	35	0	0	0	0	13007	13912	0	0	0	13007	1122	0
11	200000	2	3	2	34	11073	9787	5535	2513	1828	3731	2306	12	50	300	3738	66
12	260000	2	1	2	51	12261	21670	9966	8517	22287	13668	21818	9966	8583	22301	0	3640
13	630000	2	2	2	41	12137	6500	6500	6500	6500	2870	1000	6500	6500	6500	2870	0
14	70000	1	2	2	30	65802	67369	65701	66782	36137	36894	3200	0	3000	3000	1500	0

Table 5.3: Customer and Transactions Dataset (After Transformation)

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE_GROUP	PREV_BILL	PREV_PAYMENT	BILL	TARGET
1	20000	2	2	1	0	0	0	0	0
2	20000	2	2	1	0	0	0	0	0
3	20000	2	2	1	0	0	0	689	2
4	20000	2	2	1	0	689	689	3102	1
5	120000	2	2	2	1	3261	0	3455	1
6	120000	2	2	2	1	3455	1000	3272	1
7	120000	2	2	2	1	3272	1000	2682	1
8	120000	2	2	2	1	2682	1000	1725	1
9	90000	2	2	2	1	15549	1000	14948	1
10	90000	2	2	2	1	14948	1000	14331	1
11	90000	2	2	2	1	14331	1000	13559	1
12	90000	2	2	2	1	13559	1500	14027	1
13	50000	2	2	1	1	29547	1069	28959	1
14	50000	2	2	1	1	28959	1100	28314	1

The transformation process results in a dataset with 119,972 observations.

iii. Conversion of individual variables to required model inputs

This entailed conversion of variables such as AGE into categorical forms, creation of new variables such as the TARGET variable and conversion of categorical variables in to ‘factor’ data type.

The ‘AGE’ variable was converted to an ordinal categorical variable named ‘AGE_GROUPS’, according to the standard international age classifications provided by (United Nations New York, 1982). It was viewed that age as a factor helps to determine the behavior of cardholders based on the age groups to which they belong, and not based on their specific age in years, as discussed in Section 2.4.1.

The ‘PREV_BILL’ and ‘PREV_PAYMENT’ variables were also introduced to retain information about a customer’s payment behavior over time and the ‘TARGET’ variable was added as the dependent variable.

The ‘LIMIT_BAL’ variable was retained despite there being no information of how it changes over time (it is not normal practice for customers’ limit balance to remain constant), because its value reflects the bank’s initial assessment of the customer’s ability to repay their loan, hence it was still considered a significant predictor. For future consideration however, it would be a good improvement to the model to consider the changes to customers’ limit balances over time assuming that the information is available. For this study, the fixed value of limit balance was used for all the months per customer.

The ‘PREV_BILL’ and ‘PREV_PAYMENT’ variables reflect a customer’s current ability to repay their loan, based on past repayments. The values were obtained from the customer’s previous month’s bill and payment, and in consequence, the row corresponding to the month of April was dropped, since the dataset did not have information about the bill amount for the month of March. The row corresponding to September was also dropped since there was no information regarding the payment that the customer did for the month of September. Imputing values for the missing information was disregarded since it would lead to bias in the proportions of the ‘TARGET’ variable, which was the dependent variable.

The use of both payment and bill amounts as predictor variables and not bill only or payment only was based on the consideration that the actual implication of a single payment has to be relative to the bill amount being paid for. For example, a customer paying 100 NT Dollars against a 120 NT Dollars bill covers a bigger portion of their bill than one who paid 200 NT Dollars against a 1000 NT Dollars bill.

The ‘TARGET’ variable is a categorical variable intended to reflect the extent to which a customer will make good on their loan repayment in the current month. It is obtained by comparing the payment amount against the bill amount for a particular month. It is a nominal categorical variable whole levels are 0 (no consumption during the month), 1 (customer will revolve a balance) and 2 (customer will repay the full balance). The result is dependent on the actual amount that a customer would pay for a particular month when checked against their bill amount for that month. The ‘TARGET’ variable was the variable to be predicted such that for any month, given the customer’s sex, education, marriage status, age, bill amount, limit balance and their previous bill and payment, it was then possible to predict the state of the loan during the unobserved period. the levels 0, 1 and 2 also have a financial significance for banks since customers belonging to categories 0 and 2 do not pay any interest fees to the bank while customers who revolve balances (category 1) repay them with interest. This could be used to anticipate the revenue to be generated from issued loans. A possible improvement to the study would be to consider basing the categories of the target variable on customers’ minimum required payment amount, since where that information is available, it may lead to a more fine-grained analysis and a better way of classifying customers based on the severity of delinquency.

Figure 5.4 shows the R syntax used to achieve these transformations.

```r
#AGE
#Categorize the AGE variable (According to UN, 1981 International Standardization Rules)
#0(15-24), 1(25-44), 2(45-64), 3(65+)
data1$AGE_GROUP <- cut(data1$AGE, breaks=c(15,25,45,65,max(data1$AGE)+1),
                       include.lowest = TRUE, right = FALSE,
                       labels = c(0,1,2,3))

#drop AGE variable
data1 <- subset(data1, select = -c(AGE) )

#Convert to factors
data1$SEX <- as.factor(data1$SEX)
data1$EDUCATION <- as.factor(data1$EDUCATION)
data1$MARRIAGE <- as.factor(data1$MARRIAGE)
data1$AGE_GROUP <- as.factor(data1$AGE_GROUP)
data1$TARGET <- as.factor(data1$TARGET)
```

Figure 5.4: Algorithm for Transformation of Individual Variables

5.2.2 Multinomial Logistic Regression

The model takes in the personal characteristics of the cardholder (sex, education, marriage, and age) and the financial attributes of the cardholder (bill amount and customer rating (based on previous repayments)) and predicts the loan status for the requested period. the loan status is a categorical variable with five levels, which necessitates a multinomial logistic regression. The general equation for the logistic regression model showing the independent variables and the dependent variable is as follows:

TARGET ~ SEX + EDUCATION + MARRIAGE + AGE_GROUP + BILL + LIMIT_BAL + PREV_BILL + PREV_PAYMENT

Equation 5.1: Generalized Equation of Multinomial Logistic Regression Model

Where the 'TARGET' variable is the multilevel categorical variable to be predicted.

Since multinomial logistic regression results in generation of multiple models (*k*-1 models, where *k* is the number of levels of the dependent variable), the following models were generated, with 0 as the reference/base value for the dependent variable.

Table 5.4: Model Coefficients

```
Coefficients:
  (Intercept)       SEX2  EDUCATION1 EDUCATION2 EDUCATION3 EDUCATION4 EDUCATION5 EDUCATION6  MARRIAGE1 MARRIAGE2 MARRIAGE3 AGE_GROUP1 AGE_GROUP2 AGE_GROUP3        BILL
1   -1.319414 -0.3241911  0.46629571  0.7996683  0.7809879 -0.1270640  0.2506994 -0.3408958 0.24784773 0.4369810 0.9652863 -0.1318193 -0.1820018 -0.8204392 0.001483183
2   -1.094192 -0.1181371 -0.04764735 -0.1440935 -0.1055320 -0.2825153 -0.5260571 -1.2208515 0.02956852 0.0633042 0.4992263  0.2341087  0.2920608 -0.3570361 0.001458786
     LIMIT_BAL     PREV_BILL PREV_PAYMENT
1 -3.604350e-06 -0.0004944945 0.0005142752
2  1.826462e-06 -0.0005719541 0.0005966210

Std. Errors:
   (Intercept)        SEX2   EDUCATION1   EDUCATION2   EDUCATION3   EDUCATION4   EDUCATION5   EDUCATION6    MARRIAGE1    MARRIAGE2    MARRIAGE3   AGE_GROUP1   AGE_GROUP2
1 7.463438e-09 4.747591e-09 1.364556e-09 4.117397e-09 1.892887e-09 8.900998e-12 4.509577e-11 3.485214e-11 3.084048e-09 4.249960e-09 9.273425e-11 4.453025e-09 1.332484e-09
2 6.775019e-09 5.083636e-09 1.668200e-09 3.573778e-09 1.382936e-09 6.033509e-11 4.972697e-11 2.144507e-11 3.110912e-09 3.586559e-09 2.237546e-11 5.027387e-09 9.782989e-10
    AGE_GROUP3         BILL    LIMIT_BAL     PREV_BILL PREV_PAYMENT
1 7.531985e-12 2.177339e-05 8.189061e-08 9.757114e-06 9.823354e-06
2 7.933259e-12 2.176649e-05 6.759224e-08 9.809874e-06 9.858674e-06

Residual Deviance: 96339.5
AIC: 96411.5
```

Table 5.1 shows two sets of coefficients for the model. As is the case with multinomial logistic regression, the number of levels/classes of the dependent variables determine the number of

models that will be generated. In this case, two models were generated since the dependent variable had three levels (*k*-1 models for *k* levels of the response variable). The models performed prediction by replacing the X-value of the test instance into the equation and then obtaining the probability of the test instance belonging to either of the three classes of the response variable. The class with the highest probability was deemed to be the predicted class for that instance.

Given that the baseline/reference class is 0 (no consumption), the coefficients obtained lead to the following equations:

Let,

X_1 = Sex (Female)

X_2 = Education (Graduate School)

X_3 = Education (University)

X_4 = Education (High School)

X_5 = Education (Others 2)

X_6 = Education (Others 3)

X_7 = Education (Others 4)

X_8 = Marriage (Marriage)

X_9 = Marriage (Single)

X_{10} = Marriage (Divorce)

X_{11}= Age group (25-44)

X_{12} = Age group (45-64)

X_{13} = Age group (65+)

X_{14} = Bill

X_{15} = Limit balance

X_{16} = Previous bill

X_{17} = Previous payment

$$\log_e \frac{P(TARGET = 1(revolving\ credit))}{P(TARGET = 0(no\ consumption))} = -1.32 - 0.33X_1 + 0.47X_2 + 0.80X_3 + 0.78X_4 - 0.13X_5 + 0.25X_6 - 0.34X_7 + 0.25X_8 + 0.44X_9 + 0.97X_{10} - 0.13X_{11} - 0.18X_{12} - 0.82X_{13} + 0.001X_{14} - 0.0000036X_{15} - 0.0005X_{16} + 0.0005X_{17}$$

Equation 5.2: Equation of Logistic Regression Model (*target = 1 against 0*)

$$\log_e \frac{P(TARGET = 2(full\ payment))}{P(TARGET = 0(no\ consumption))} = -1.09 - 0.12X_1 - 0.05X_2 - 0.14X_3 - 0.11X_4 - 0.28X_5 - 0.53X_6 - 1.22X_7 + 0.03X_8 + 0.06X_9 + 0.49X_{10} + 0.23X_{11} + 0.29X_{12} - 0.36X_{13} + 0.001X_{14} + 0.0000018X_{15} - 0.0006X_{16} + 0.0006X_{17}$$

Equation 5.3: Equation of Logistic Regression Model (*target = 2 against 0*)

The model equations can be interpreted as follows:

i. A one-unit increase in the limit balance (X_{15}) is associated with the decrease in the log-odds of being in the revolving credit class versus no consumption class in the amount of 0. 0000036

ii. A one-unit increase in the limit balance (X_{15}) is associated with the increase in the log-odds of being in the full payment class versus no consumption class in the amount of 0. 0000018
iii. The log-odds of being in the revolving credit class versus being in the no consumption class will decrease by 0.33 if the gender moves from male to female
iv. The log-odds of being in the full payment class versus being in the no consumption class decreases by 0.12 if the gender moves from male to female.

The same interpretation can be done for all the other variables.

To obtain the ratio of the probability of being in one category over the probability of being in the reference category (the odds) and not the log-odds, the equation on the right-hand side is exponentiated, hence the exponentiated regression coefficients are the odds for a unit change in the predictor variable (UCLA: Statistical Consulting Group, n.d.).

Selection of predictor variables to retain depended on the p-value score for each coefficient. P-value tests the null-hypothesis that the coefficient of a predictor is equal to zero, meaning that the predictor has no effect on the outcome that is observed. A p-value that is less than 0.05 means that the null hypothesis can be rejected, hence the predictor has significance in the model (Frost, n.d.). In this study, since the library used did not indicate the p-value calculation for the coefficients, the value was calculated using Wald tests (z hypothesis test). Table 5.2 shows the z-values and Table 5.3 shows the p-values for the regression coefficients.

Table 5.5: Coefficients' Z-test Result

	(Intercept)	SEX2	EDUCATION1	EDUCATION2	EDUCATION3	EDUCATION4	EDUCATION5	EDUCATION6	MARRIAGE1	MARRIAGE2	MARRIAGE3	AGE_GROUP1	AGE_GROUP2	AGE_GROUP3
1	-176783681	-68285385	341719688	194216952	412590812	-14275249842	5559267088	-9781201949	80364432	102820032	10409167131	-29602195	-136588302	-108927362526
2	-161503888	-23238705	-28562133	-40319644	-76310121	-4682437017	-10578908027	-56929236895	9504778	17650400	22311332995	46566683	298539410	-45004978791

	BILL	LIMIT_BAL	PREV_BILL	PREV_PAYMENT
1	68.11906	-44.01421	-50.68041	52.35231
2	67.01978	27.02176	-58.30392	60.51736

Table 5.6: Coefficients' p-values

	(Intercept)	SEX2	EDUCATION1	EDUCATION2	EDUCATION3	EDUCATION4	EDUCATION5	EDUCATION6	MARRIAGE1	MARRIAGE2	MARRIAGE3	AGE_GROUP1	AGE_GROUP2	AGE_GROUP3	BILL	LIMIT_BAL
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	PREV_BILL	PREV_PAYMENT
1	0	0
2	0	0

The results shown in Table 5.3 show that all predictors were statistically significant since their p-values were all below 0.05 (all 0).

In the model, indicator variables for the categorical variables were created to indicate the presence or absence of the categorical effect. This is important because the levels of the categorical variables do not intrinsically contain any statistical significance. For example, the value '2' in the 'EDUCATION' variable does not in itself have weight in determining the outcome of the dependent variable as compared to the value '1'. This is because in nominal

variables, the values are expected to be mathematically equal. As a result, indicator variables are created which are binary in nature (value 0 or 1). This explains the presence of the 'EDUCATION1', 'EDUCATION2', 'EDUCATION3' and other variables in the model.

5.3 System Implementation

The system provided an implementation of high level user functionality and interfaced with the default prediction model to perform prediction. A web application provided the user interface for the loan officer to upload credit cards data and to obtain predictions of loans. For model fitting and prediction, the Java-based web system would interface with an R application through a Java-R bridge that enabled running on R code on a Java environment. The system used a MySQL database to store details of users, their user groups and permissions. The following technology was used in the development environment for the system:

i. Java Spring Boot Framework – for development of the web application
ii. MySQL database
iii. rJava – an R-to-Java interface

Software flow for the system includes a user authentication process, where the loan officer or system administrator provides their login credentials after which they access their portals whose functionality differs based on the permissions that they are assigned. After successful authentication, the system administrator accesses their profile from where they can perform system configuration functions such as creation of users and removal of users. A loan officer's profile page enables them to upload a credit card dataset for model training and to enter cardholder details for new customers or to search for existing customers and obtain a prediction for their existing loans.

Figure 5.5 shows the data upload page from which the loan officer trains the model.

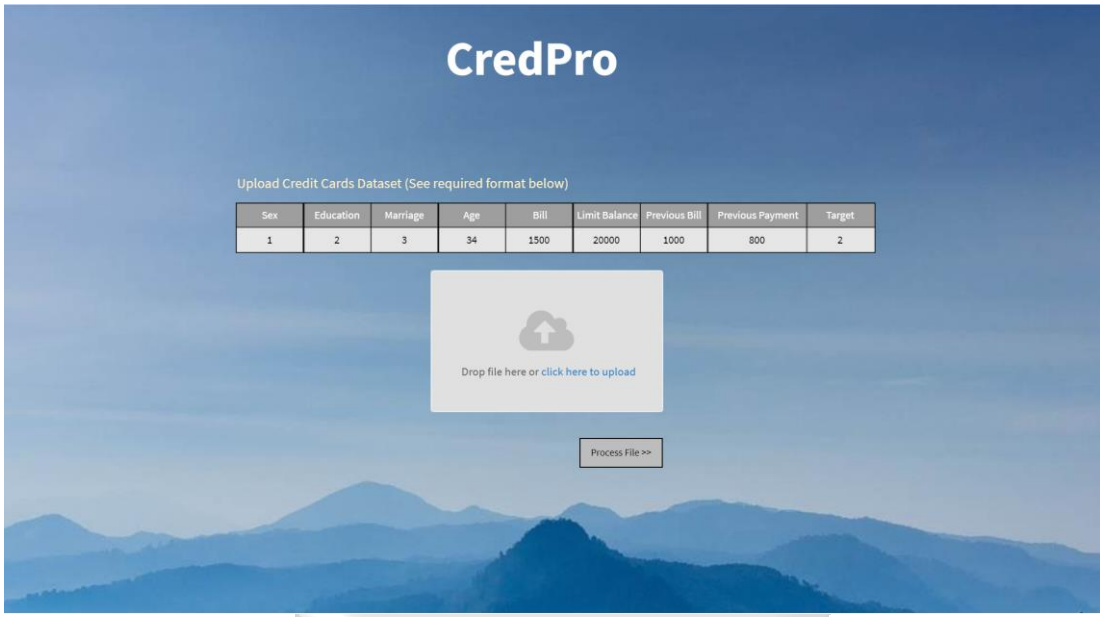

Figure 5.5: ‘Model Training’ Interface

Figure 5.6 shows the loan officer’s profile from which they can get predictions for customer’s loans.

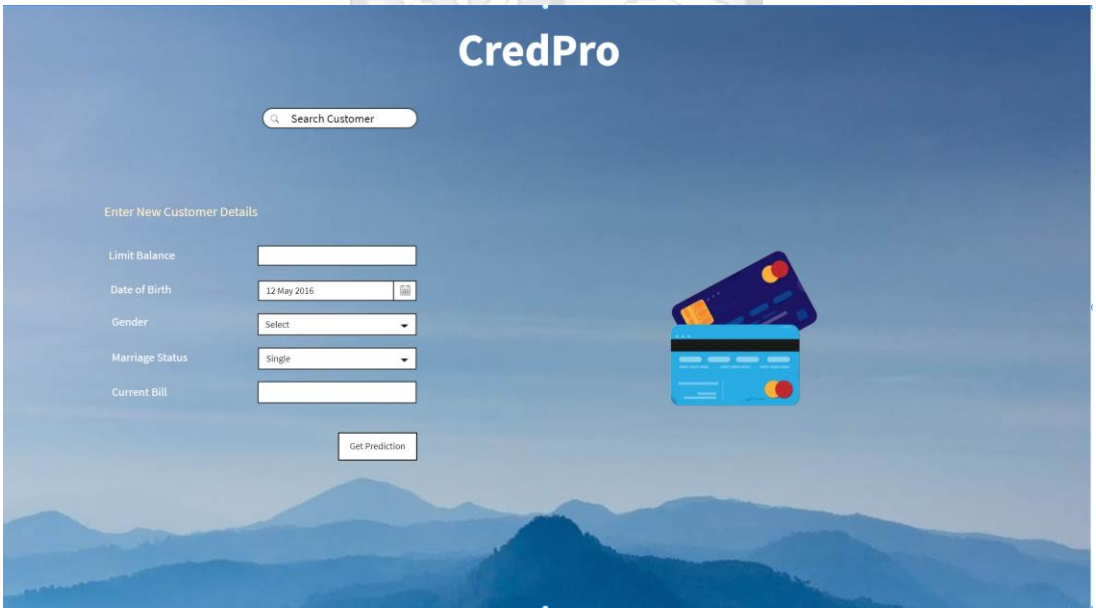

Figure 5.6: ‘Get Prediction’ Interface

5.4 Testing and Validation

Testing and validation efforts were done with the aim of ensuring that the model could predict the loan status of a cardholder given their personal attributes and financial attributes and give

reliable results which could be used in decision making processes of financial institutions aimed at managing the credit risk or profitability. Reliability of results was measured using the parameters of model accuracy, precision and recall.

A 10-fold cross validation was performed to validate the model. The model reached convergence after 60 iterations in all 10 validation tests, showing that it was a good fit for the data used for testing. The mean accuracy of predictions of loan statuses for customers during the five-months' duration was found to be 85.28%, with a mean recall of 83.71% and a mean model precision of 77.87%.

According to the results obtained, it can be seen that the model outperformed Leow and Crook's (2014) model which had an overall loan status prediction accuracy of 83% in predicting customer loan statuses during the repayment periods. Comparison was made with this study since this is the only other study which considers multi-state outcomes of loans during the periods leading up to default, and not only the default event itself. To be precise, the model developed by Leow and Crook (2014) was able to predict loan delinquency by clients during the months leading up to default, and could also predict the default event as well. This is in contrast to other classification models such as Islam (2018), who was able to attain a 95.4% classification accuracy using extremely randomized trees but whose dependent variable was the binary outcome: default or non-default with no consideration of outcomes in between the periods of observation. It is therefore reasonable to do comparison with the model developed by Leow and Crook (2014) only due to a similar contextualization of the problem at hand.

Chapter 6: Discussion

6.1 Introduction

This study led to the development of a model for prediction of loan defaulting by credit card customers using logistic regression analysis. The data used in the study contained personal attributes of cardholders and their borrowing and repayment histories for a period of six months. By using the model, it was possible to predict the loan status of a customer given their personal attributes (gender, education level, marriage status and age group) and their financial attributes (credit limit, previous bill, previous payment and current bill amount) with an accuracy of 85.28%.

6.2 Discussion

From the analysis performed, it was found that the most important variables in determining the outcome of a loan in their order are as shown in Table 6.1. The measure of variable importance is the sum of absolute values of coefficients on the generated models, a metric used in varImp() method to measure variable importance in a regression model.

Table 6.1: Variable Importance

	Variable	$\sum$**|coefficients|**
1	EDUCATION6	1.561747
2	MARRIAGE3	1.464513
3	AGE_GROUP3	1.177475
4	EDUCATION2	0.9437617
5	EDUCATION3	0.8865199
6	EDUCATION5	0.7767565
7	EDUCATION1	0.5139431
8	MARRIAGE2	0.5002852
9	AGE_GROUP2	0.4740626
10	SEX2	0.4423282
11	EDUCATION4	0.4095792
12	AGE_GROUP1	0.3659280
13	MARRIAGE1	0.2774163
14	BILL	0.002941968
15	PREV_PAYMENT	0.001110896
16	PREV_BILL	0.001066449
17	LIMIT_BAL	0.000005430812

From the table, it is evident that a cardholder's personal characteristics played a more significant role in determining the resultant loan status than their transactional activities.

6.2.1 Cardholder Characteristics and their Influence on Defaulting

The study focused on four attributes that fall in the category of cardholders' personal attributes: gender, education level, marriage status and age. These form a cardholder's socio-demographic factors and as can be seen from the results obtained, they play a significant role in explaining the cardholders' behavior with regards to repayment of debts. This collaborates the findings of Marjo (2010) and Kocenda and Vojtek, (2009) as discussed in Section 2.4.1. Kanyi (2009) in his study found that gender did not influence loan defaulting for credit cards, which differed with the findings of this study in which gender is shown to be a significant predictor for default, although not the most significant but nonetheless could not be disregarded.

Table 6.1 indicates that the socio-demographic factors are more significant as predictors of loan default as compared to the transactional activities of the customer. This revealed that these factors were the drivers which influence a customer's behavior, and since transactional activities can themselves be regarded as part of customer behavior, this may therefore be considered a credible explanation to the findings.

6.2.2 Transaction Activities and their Influence on Defaulting

Transactional activities were reflected in the bill amount, previous payment, previous bill and credit limit variables. Credit limit only indirectly relates to the transaction activities since the value is adjusted depending on the customer's credit card usage. It reveals the customer's creditworthiness from the bank's perspective depending on how the customer transacts and hence falls into the category of variables related to transaction activities.

All variables related to a customer's transaction activities were retained as predictor variables for loan defaulting since they were all found to be statistically significant for the model. This agrees with the findings of Lee et al. (2011) that a customer's transaction activities determine whether they will eventually make good on their loan or not, and they therefore may be used as predictors for future default tendencies.

In comparison to the personal attributes of cardholders, the findings of this study indicate that transaction activities are less significant variables in determination of loan defaults. Table 6.1 shows that personal attribute variables all come before variables related to transaction activities with regards to significance as predictors.

6.2.3 Prediction of Defaulting using Logistic Regression

Logistic regression is used in modeling the relationship between one or more independent variables and a categorical dependent variable through a linear function (Agbemava et al.,

2016). In this study, multinomial logistic regression was successfully used to model the relationship between attributes of credit cardholders and their loan repayment behavior for given time periods. The model was set up with eight input variables, four being categorical and four being non-categorical, and a three-level categorical dependent variable. Prediction of resultant loan statuses was based on determination of probabilities of the loans belonging to either levels of the dependent variable, and then selecting the level with the highest probability and declaring that as the predicted level.

Logistic regression was appropriate for this study since it was possible to model the relationship between categorical and non-categorical independent variables with a categorical dependent variable and at the same time obtain a two-fold advantage from the analysis: to determine which independent variables were significant for the analysis and also to use those variables to predict loan defaulting.

6.2.4 Performance of the Developed Model in Predicting Defaults

The model achieved 85.28% prediction accuracy for customer's resultant loan states given their personal attributes and transaction activities. This was an acceptable result and the model could be extended to predict default even in different forms of definitions for the default event. For instance, one bank may deem default to have occurred based on the performance of its debtors over a period of six months, stating that a customer is a defaulter if they have not repaid at least 70% of their loans and interest charges as at the end of those six months. Another bank may have different criteria for designating customers as defaulters or non-defaulters. The model developed during this study can be extended to predict default events no-matter how the conditions necessary for one to be declared a defaulter have been configured.

6.2.5 Scientific Contribution

The model developed provided the prediction of a customer's loan status given their personal characteristics and transactions history, and for any unobserved period (month), the model could give the proportion of a bank's cardholders who would have no balance, those who would revolve a balance and those who would repay their loans in full. This presents a unique way of looking at a bank's risk portfolio in that the focus is not only of the eventual default event, but also on customers' borrowing and repayment behaviors during the times leading up to default. This approach has been utilized by Leow and Crook (2014) who developed a model based on intensity modeling of credit risk using a hazard function, leading to formation of a matrix of transition probabilities between loan states within the periods leading up to the eventual default

event. This study uses logistic regression to achieve period-by-period (month-by-month) predictions of loan states without creation of a matrix of transition probabilities and achieving better performance in prediction.

Chapter 7: Conclusion and Recommendations

7.1 Conclusion

This research was centered around the attainment of five objectives with the overall goal of developing a credit card loan default prediction model. The five objectives are as follows:

(i) To analyze the cardholder characteristics that influence credit card loan defaulting

It can be concluded that cardholders' personal attributes greatly influence their tendency to default. This was supported both by the literature that was reviewed and also by the findings of this study themselves.

(ii) To analyze the transaction activities between customers and their banks that influence credit card loan defaulting

An analysis of how cardholders use their credit cards with regards to borrowing and repayments may be used as predictors of delinquency and hence are important factors to consider in the management of risk portfolios of financial institutions. Of importance to note is that while factors related to transaction activities of customers are less powerful predictors of eventual loan status as compared to the cardholder characteristics, information about them builds up over time and is more dynamic in nature relative to the personal characteristics. This means that given sufficient time, and with enough data regarding the trends of these variables, it will be possible to obtain better performance on default prediction models.

(iii) To examine the methods and technologies that are used to predict loan defaulting for credit card accounts

In this study, multinomial logistic regression analysis was used to develop a model for prediction of loan defaulting with a success rate of 85.28%. It can be concluded that this method is a good fit for similar contexts and depending on the problem area, machine learning approaches are worthwhile considerations for harnessing the value of data towards generation of feedback, forecasts and inputs towards decision-making processes of organizations.

(iv) To develop a default prediction model for credit card accounts

This study led to the successful development of a model for prediction of default and a system that implements that model to provide users with an interface to use the prediction functionality of the model.

(v) To test the performance of the model in predicting credit card loan defaulting

The model performance was tested based on the parameters of accuracy, recall and precision and compared to a similar model developed from a different study thereby showing better performance. An accuracy of 85.28% was attained, a recall level of 83.71% and a precision of 77.87%.

7.2 Recommendations

From the findings of the study, the following recommendations can be made:

i. Since it is evident that cardholders' personal characteristics and their transaction histories are significant predictors for default, it is recommended that financial institutions base the management process of their loan portfolios on these factors and create customer profiles on this basis
ii. There is a need for financial institutions to adopt default prediction models to forecast the performance of their customers with regards to issued loans since the models are efficient and cost effective.
iii. In order to obtain better performance on credit risk models, it is recommended that data regarding the cardholders' minimum required repayment amount and changes of customers' credit limits over time be factored in to the modeling process.

7.3 Future Work

An addition to the model would be to consider applying time-varying macroeconomic variables while performing the analysis. Macro-economic variables affect all customers but different customers respond differently to different levels of these variables and a time-variant consideration changes the responses when the analysis is performed at different time periods. In addition, basing the categories of the dependent variable on the customers' minimum required repayment amount would also be a valuable modification of this study since this basis has more fine-grained financial implications in terms of revenue gains/losses that would be forecasted by the model in unobserved periods.

References

Agbemava, E, Nyarko I. K., Adade, T. C., & Bediako, A. K. (2016, January). Logistic Regression Analysis of Predictors of Loan Defaults by Customers of Non-Traditional Banks in Ghana. *European Scientific Journal* 12(1), 175-189.

Amadeo, K. (2019, April 17). *Average US Credit Card Debt Statistic*. Retrieved April 18, 2019, from https://www.thebalance.com/average-credit-card-debt-u-s-statistics-3305919

Bacham, D., & Zhao, J. (2017, July). *Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling*. Retrieved April 21, 2019, from https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling

Bellotti, A., & Crook, J. (2009). Forecasting and Stress Testing Credit Card Default using Dynamic Models. International Journal of Forecasting. 29. 1-36. 10.1016/j.ijforecast.2013.04.003.

Benson, C. C., & Loftesness, S. (2014, February 17). *Payments Systems in the U.S.* (2nd ed.). California: Glenbrook Press.

Bronshtein, A. (2017, April 11). *A Quick Introduction to K-Nearest Neighbors algorithm.* Retrieved April 21, 2019, from https://blog.usejournal.com/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

Brown, K., & Moles, P. (2014). *Credit Risk Management.* Retrieved February 27, 2020, from http://www.ebsglobal.net/EBS/media/EBS/PDFs/Credit-Risk-Management.pdf

Chen, S., Hardle, W. K., & Moro, R. A. (2011). Modeling Default Risk with Support Vector Machines. *Quantitative Finance.* 11(1), 135-154

Credit Risk Ratings. (2008, May). Retrieved April 21, 2019, from https://www.dico.com/design/4_17_P3_Eng.html

Donges, N. (2018, February 22). *The Random Forest Algorithm*. Retrieved April 21, 2019, from https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Dormehl, L. (2019, January 5). *What is an artificial neural network? Here's everything you need to know*. Retrieved April 21, 2019, from https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/

Drugov, V. G. (n.d.). Default Payments of Credit Card Clients in Taiwan from 2005. Retrieved April 23, 2019, from https://rstudio-pubs-static.s3.amazonaws.com/281390_8a4ea1f1d23043479814ec4a38dbbfd9.html

Dwyer, B. (2018, March 28). *Credit Card Processing: How it Works*. Retrieved May 05, 2019, from https://www.cardfellow.com/blog/how-credit-card-processing-works/

Eriksson, U. (2012, April 05). Functional vs Non-Functional Requirements. Retrieved January 23, 2020, from https://reqtest.com/requirements-blog/functional-vs-non-functional-requirements/

Evans, S. D., & Schmalensee, R. (2005). *Paying with Plastic: The Digital Revolution in Buying and Borrowing* (2nd ed.). London: The MIT Press.

Federal Reserve Bank of St. Louis (2019, February 20). *Delinquency Rate on Credit Card Loans, All Commercial Banks*. Retrieved April 18, 2019, from https://fred.stlouisfed.org/series/DRCCLACBN

Frost, J. (n.d.). *How to Interpret P-values and Coefficients in Regression Analysis*. Retrieved March 19, 2020, from https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

Gandhi, R. (2018). *Support Vector Machines – Introduction to Machine Learning Algorithms*. Retrieved April 22, 2019, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Geurts, P., Ernst, D. & Wehenkel, L. (2006). Extremely randomized trees. 63: 3. https://doi.org/10.1007/s10994-006-6226-1

Gonçalves, L. (2019, May). *What is Agile Methodology*. Retrieved July 14, 2019, from https://luis-goncalves.com/what-is-agile-methodology/

Gross, D. B. & Souleles, N. S., "An Empirical Analysis of Personal Bankruptcy and Delinquency," Review of Financial Studies, Vol. 15, No. 1, 2002, pp. 319-347. doi:10.1093/rfs/15.1.319

Gurusamy, S. (2009). *Merchant Banking and Financial Services* (3rd ed.). India: Tata McGraw-Hill Education

Hadzikadic, M. (2002, November 24). *Credit Analysis of Czech Bank*. Retrieved April 22, 2019, from https://webpages.uncc.edu/mirsad/itcs6265/group1/index.html

Husejinovic, A., Kečo, D., & Mašetić, Z. (2018). Application of Machine Learning Algorithms in Credit Card Default Payment Prediction. *International Journal of Scientific Research.* 7. 425. 10.15373/22778179#husejinovic

Islam, S. R. (2018). *An efficient technique for mining bad credit accounts from both olap and oltp.* Ph.D. dissertation, Tennessee Technological University.

Islam, S. R., Eberle, W., & Khaled, G. S. (2017). *Mining Bad Credit Card Accounts from OLAP and OLTP*. 129-137. 10.1145/3093241.3093279

Kanyi, K. F. (2015, October). *Influence of socio-demographic, behavioral and economic determinants on credit cards default in commercial banks in kenya.* PhD Thesis. Kenyatta University. Retrieved February 18, 2020, from https://ir-library.ku.ac.ke/

Kenton, W. (2019, April 8). *Industrial Production Index (IPI).* Retrieved April 20, 2019, from https://www.investopedia.com/terms/i/ipi.asp

Kiarie, F. K., Nzuki, D. M., & Gichuhi, A. W. (2015). Influence of Socio-Demographic Determinants on Credit Cards Default Risk in Commercial Banks in Kenya. *International Journal of Science and Research.* 4(5), 1611-1615

Krichene, A. (2017, February 2). Using a naïve Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank. *Journal of Economics, Finance and Administrative Science*. 22(42), 3-24

Kumar, M., Goel, V., Jain, T., Singhal, S., & Goel, L. M. (2018, April). Neural Network Approach to Loan Default Prediction. *International Research Journal of Engineering and Technology (IRJET).* 5(4), 4231-4234

Lee, C., Lin, T. T., & Chen, Y. (2011). An Empirical Analysis of Credit Card Customers' Overdue Risks for Medium-and Small-Sized Commercial Bank in Taiwan. *Journal of Service Science and Management*. 4. 234-241. 10.4236/jssm.2011.42028.

Leow, M., & Crook, J. (2014). Intensity models and transition probabilities for credit card loan delinquencies. *European Journal of Operational Research.* 236 (2014) 685–694

Leow, M., & Crook, J. (2015). *Exploring the Effects of Macroeconomic Variables on Credit Card Delinquency and Default Behaviour.* Working paper, Credit Research Centre, University of Edinburgh

Ma, Y.H. (2020) Prediction of Default Probability of Credit-Card Bills. *Open Journal of Business and Management*, 8, 231-244. https://doi.org/10.4236/ojbm.2020.81014

Obare, D. M. & Muraya, M. M. (2018). Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults. *American Journal of Applied Mathematics and Statistics*. 6(6), 266-271

Ochieng, J. (2019, February 18). *Interest Rate Cap Two Years On: Outcomes for Kenya's Economy*. Retrieved April 20, 2019, from http://kippra.or.ke/interest-rate-cap-two-years-on-outcomes-for-the-kenyan-economy/

Oula, B. N. (2013, July). *Exploratory loan data analysis & modelling time to default using survival analysis techniques.* Master's thesis. University of Nairobi. Retrieved November 10, 2019, from http://erepository.uonbi.ac.ke/

Patel, S. (2017, May 3). *Machine learning 101*. Retrieved April 23, 2019, from https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

Pershad, R. (2000). *A Bayesian Belief Network for Corporate Risk Assessment.* Master's thesis. University of Toronto. Retrieved February 14, 2020 from https://tspace.library.utoronto.ca

Philemon. (2018, June 5). *Managing Credit Risk: An Effective Approach with FinTech Data Science.* Retrieved April 20, 2019, from https://medium.com/@comeblossom.gh/managing-credit-risk-an-effective-approach-with-fintech-data-science-f8303dcee360

Radu, C. (2003). *Implementing Electronic Card Payment Systems*. Norwood, MA: Artech House.

Rajani, V. (2009). *An evaluation of business deals using plastic money in Kerala.* Ph.D. dissertation, Cochin University of Science and Technology.

Reid, C. (2019, January 22). *An Overview of Credit Card Processing Transaction Types*. Retrieved June 16, 2019, from https://www.helcim.com/article/overview-credit-card-transaction-types/.

Research to Action (2015, February 13). *Research quality and Think Tanks: definition, responsibility and impact*. Retrieved April 21, 2019, from https://www.researchtoaction.org/2015/02/research-quality-responsibility-impact-role-think-tanks/

Rokad, B. (2019, August 1). *Machine Learning Approaches and its Application*. Retrieved March 10, 2020, from https://medium.com/datadriveninvestor/machine-learning-approaches-and-its-applications-7bfbe782f4a8

Saxena, S. (2018, May 11). *Precision vs Recall*. Retrieved July 27, 2019, from https://towardsdatascience.com/precision-vs-recall-386cf9f89488

Shain, S. (2019, March 18). *How does a credit card work? Here's an Easy-to-Understand Guide and Credit Card Definition*. Retrieved April 18, 2019, from https://www.creditcardinsider.com/blog/how-does-a-credit-card-work/

Spacey, J. (2017, August 12). *11 Examples of Usability Requirements.* Retrieved January 23, 2020, from https://simplicable.com/new/usability-requirements

Steele, J. (2019, March 19). *Best premium credit card for 2019*. Retrieved July 14, 2019, from https://www.thesimpledollar.com/credit-cards/best-premium-credit-card/

UCLA: Statistical Consulting Group (n.d.). Multinomial logistic regression | R data analysis examples. Retrieved March 19, 2020, from https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

Williams, R. (2019, February 7). *Multinomial Logit Models – Overview*. Retrieved March 9, 2020, from https://www3.nd.edu/~rwilliam/stats3/Mlogit1.pdf

Yüksel, S., Zengin, S., & Kartal, M. T. (2017). Identifying the Macroeconomic Factors Influencing Credit Card Usage in Turkey by Using MARS Method. China-USA Business Review. 15. 611-615. 10.17265/1537-1514/2016.12.003.

Appendix

Appendix A: Strathmore University Ethical Approval Certificate

Strathmore UNIVERSITY

14th January 2020

Mr Munene, Lewis
munene.lewis@strathmore.edu

Dear Mr Munene,

RE: A Model for Predicting Credit Card Loan Defaulting using Account Transaction Activities

This is to inform you that SU-IERC has reviewed and **approved** your above research proposal. Your application approval number is **SU-IERC0606/19** .The approval period is **14th January, 2020 to 13th January, 2021.**

Appendix B: Research License (NACOSTI)

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **847770**

Date of Issue: **10/March/2020**

RESEARCH LICENSE

This is to Certify that Mr.. Lewis Munene Muchiri of Strathmore University, has been licensed to conduct research in Nairobi on the topic: A Model for Predicting Credit Card Loan Defaulting using Account Transaction Activities for the period ending : 10/March/2021.

License No: **NACOSTI/P/20/3917** Ammended

847770

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

Appendix C: Originality Report