2021

A Model for predicting pre-delinquency of credit card accounts using Extreme Gradient boosting.

Kisengese, Antony Mwawuganga
Faculty of Information Technology
Strathmore University

A Model for Predicting Pre-Delinquency of Credit Card Accounts using Extreme Gradient Boosting

By

Kisengese, Antony Mwawuganga

Submitted to the School of Computing and Engineering Sciences in partial fulfillment of the requirement of the Degree of Master of Science in Information Technology

Strathmore University

September 2021

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Kisengese, Antony Mwawuganga

07 September 2021

Approval

The thesis of **Kisengese, Antony Mwawuganga** was reviewed and approved by the following:

Dr. Vincent Omwenga
Research Director and Senior Lecturer, School of Computing and Engineering Sciences
Strathmore University

Dr. Joseph Orero
Dean, Faculty of Information Technology
Strathmore University

Dr. Bernard Shibwabo
Director of Graduate Studies

List of Abbreviations

ATM	Automated-Teller Machine
CSV	Comma-Separated Values
CVV	Card Verification Value
ISO	International Organization for Standardization
MCC	Merchant Category Code
OTP	One-Time Password
PAN	Primary Account Number
PSI	Population Stability Index
SVM	Support Vector Machines
TPB	Theory of Planned Behavior
WoE	Weight of Evidence

Abstract

Credit risk is one of the significant risks that financial institutions that advance credit in credit cards are exposed to. Credit card accounts are usually classified as "good" or "bad" depending on the propensity of the cardholder to settle their debt on time. The latter usually pose a significant negative impact to the issuer's books when the credit card account falls into late collections and recoveries are futile resulting to bad debts. Ensemble classifier algorithms have demonstrated greater performance in classification and regression problems due to their ability to trade-off bias and variance factors. In this study, an Extreme Gradient Boosting ensemble classifier was implemented based on cardholder personal characteristics and transaction patterns with the aim to minimize defaults in the late collection stages by identifying credit card accounts that exhibit early signs of delinquency way before the cardholder misses payments. A credit card dataset from the UCI Machine Learning repository was used to train and validate the model, which achieved a prediction accuracy of 81.62% and outperformed a set of single classifiers that were used in benchmarking. Depending on each score, the issuer will make informed decisions of how well to proactively engage the cardholder to identify the best way of intervening in their financial situation and mitigate the risk of missing payments.

Keywords: Credit risk, Credit Scoring, Delinquency, Extreme Gradient boosting

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1: Introduction

1.1 Background to the study

Credit cards are payment instruments issued by financial institutions to their customers for accessing credit limits for making payments, after which customers repay the issuer sometime in the future based on a credit agreement. This credit limit is like a loan. The card is a physical representation of a financial account held with the issuer. Therefore, all transactions performed using the card are debited against this account. Every card transaction reduces the available credit with the same transaction amount for which the cardholder can access future payments. At the end of every month, the cardholder's total debt is the amount spent in that month, plus the interests accrued or any other agreed fees. Cash withdrawals usually attract high transaction fees and interests to discourage customers from doing so since this is not the primary use of these facilities. Besides, a credit card offers a cardholder great payment convenience with various benefits including but not limited to the ease of performing cashless transactions, the ability to track all expenses at once, and offering an accurate documentation history of one's creditworthiness (Lin et al., 2019; Wong and Lynn, 2019).

A cardholder is usually required to make the minimum payment agreed on their total card debt at the end of every month, but also with the liberty of paying the whole debt owed. When one fails to fulfill this credit agreement, the card is said to be delinquent (Finlay, 2010). A preliminary assessment is usually taken by the issuer to determine the account's status, if the account should fall into collections to recover the debt at the earliest opportunity (Finlay, 2010) or other treatment strategies are to be used. This assessment is undertaken to ascertain whether a failure to submit payments is due to fraud or a cardholder is under financial stress. When the delinquency period is exhausted, and the cardholder fails to make payments, the account falls into collections in which an issuer follows recovery strategies that are part of their operations, from

internal debt collection and recovery teams to acquiring services of debt recovery agencies, all in a bid to motivate the cardholder to repay their debt. When repayment is impossible, the debt is written off, and the account becomes default, which is a terminal state in the credit card life cycle.

Pre-delinquency is the state before the credit cardholder misses payments. Before the expected minimum payment is made, this time frame offers the issuer an opportunity to spot any form of financial duress that may result in delinquency. Finlay (2010, p.140) argues that cardholders usually miss payments when their financial situations have changed, such as loss of income, which consequently changes their credit usage patterns, and these patterns can be detected with the use of a pre-delinquency scoring model. A high pre-delinquency score defines a high propensity of the cardholder missing a payment, while a low score the vice versa. Depending on the score, the issuer can make informed decisions about whether to contact customers and take pre-emptive interventions that may prevent the customer from becoming delinquent. For example, since accounts with low pre-delinquency show more significant self-cure signs, the issuer can send informative emails or messages to the cardholder, which reminds them of upcoming bills and has the likelihood of triggering those customers who don't miss payments to repay on time. On the other hand, accounts with high pre-delinquency scores may require the issuer to contact the cardholder to determine any mitigating actions that can be undertaken, such as suspending repayments for a while for the customer to recover their financial footing (Finlay, 2010). Selecting the proper strategy, message, and channel to approach a cardholder provides the issuer a far better likelihood of securing repayments whereas maintaining a positive relationship with them (Esgalhado et al., 2019).

Pre-delinquency scoring is a relatively new scoring model applied to existing customers to identify those up-to-date customers who are likely to miss their next payment (Finlay, 2010). This scoring model differs from credit scoring models in that the goal of pre-

delinquency scoring is not to decide about further lending, but rather to inform the decision of whether to contact customers and take pre-emptive actions that may avoid the customer from becoming delinquent. However, credit scoring techniques can also be applied in pre-delinquency scoring as they are both concerned with predicting the probability of undesirable customer behavior in the future (Lessmann et al., 2015).

Multiple studies have been undertaken to develop scoring models, with logistic regression so far being the de-facto technique that has been used extensively (Onay, & Öztürk, 2018). Lessmann et al. (2015) performed a benchmark of 41 classification algorithms and observed that ensemble classifiers had a better predictive performance compared to single classifiers such as support vector machines, artificial neural networks. This study will use an extreme gradient boosting (EGBoost) ensemble to fit and build the delinquency scoring model.

1.2 Problem Statement

Credit risk management is one of the management practices that financial institutions involved in consumer lending must undertake and has shifted from being a necessary business evil to a strategic survival imperative. When compared to other risks such as market, operational and liquidity risks, credit risk is the biggest and constitutes about 50% of the total risk exposure. Canner and Elliehausen (2013) conducted research to examine consumer behavior, experiences, and attitudes with regard to credit card debt in the effect of changing economic conditions and regulatory changes, and observed that, 30% of credit cardholders fell behind on payments at least once, 20% fell being at least 30 days, and 80% of credit cardholders sometimes or hardly ever paid their monthly balances in full. According to Finlay (2010), credit card debts that are written-off due to credit cardholders failing to repay accounts for about one third of all costs in consumer credit management and rise considerably in downturn economic conditions. The perpetual risk that issuers face and that must continue to be addressed is the likelihood of credit cardholders borrowing more than can they are willing to repay.

Issuers have developed and implemented credit scoring models to manage risky customers ex ante by modeling a customer's creditworthiness into either a "good" or "bad" category (Teng & Lee, 2019; Chou & Lo, 2018). This is applied to both new and existing customers. For new customers, the goal is to access the customer's potential risk once they have been granted a credit facility, while for existing customers, the goal is to forecast their future behavior over a given time horizon (Onay & Öztürk, 2018). In the latter, behavioral scoring, the customer's risk is measured, and their value assigned given their actual spending and repayment behavior. Whilst these models being extensively used, the categorization of customers in into two groups, either good or bad fails to account for the customer's risk value that can inform the issuer on the propensity of missing a payment for the customer in question.

Pre-delinquency scoring therefore enables the issuer to evaluate the value at risk way earlier in the collections cycle. Therefore, more efforts can be channeled to customers who show non-self-cure signs using a proactive approach to intervene in their financial situations using personalized arrangements. For example, in a case when the credit card account has a high pre-delinquency score, and there exists some headroom between the customer's balance and their credit limit, the issuer can lower the credit limit to mitigate the risk of exposure if the account was to fall into delinquency (Finlay, 2010). Other interventions include and are not limited to account closures, payment reminder notifications, and change in the minimum repayment amount.

This research proposes a prediction model that considers the personal characteristics of the cardholder, their credit usage, and payment behavior to predict the likelihood of missing payments in each billing cycle, and possibly in upcoming billing cycles.

1.3 Objectives

1.3.1 General Objective

The study aimed to develop a prediction model based on cardholder characteristics and transaction activities to minimize the number of credit card accounts that fall into delinquency

1.3.2 Specific Objectives

i. To analyze the cardholder characteristics and transaction activities that influence pre-delinquency credit card accounts
ii. To examine the classifier algorithms used in predicting pre-delinquency of credit card accounts
iii. To develop a model for predicting pre-delinquency of credit card accounts using ensemble classifier algorithms
iv. To test the performance of the model in predicting the pre-delinquency of credit card accounts

1.4 Research Questions

1. How do cardholder characteristics and transaction activities influence pre-delinquency of credit card accounts?
2. What are the classifier algorithms used to predict pre-delinquency of credit card accounts?
3. How can the model for predicting pre-delinquent credit card accounts based on ensemble classifier algorithms be developed?
4. How can the performance of the developed model be tested?

1.5 Justification

Credit card debt has been rising as more and more credit accounts fall into delinquency with revolving balances. Prediction models come handy to classify customers into segments for more targeted interventions depending on the value at risk and customer profile. This can lead to increased issuer-cardholder connection and raise the chances of cardholders paying, and effective settlement approaches can also be designed especially for cardholders under financial duress. Therefore, the issuer can position themselves to expand on relationships demonstrating that they care about cardholders' financial health and well-being. Also, this reduces strain on collections operations by keeping low-risk and self-cure accounts out of delinquency.

1.6 Scope and Limitation

The research's scope was limited to learning from publicly available data related to credit card accounts with has limited demographic, socio-economic, credit loan, and payment history information. This is due to the sensitivity of the credit card data which banks treat as confidential and private. The study will also be limited to predicting the possibility of a credit cardholder missing a payment based on the patterns established in the data.

Chapter 2: Literature Review

2.1 Introduction

This chapter will cover relevant literature related to the study. It will investigate the theory of planned behavior to construct the independent variables that influence behavioral scoring in the study. The empirical review will provide both statistical and machine learning techniques that have been used in credit scoring.

2.2 Theory of Planned Behavior (TPB) in Predicting Human Behavior

Icek Ajzen postulated the theory of planned behavior in 1985 as a model to predict and understand human behavior in that behaviors are immediately determined by behavioral intentions and, under certain circumstances, perceived behavioral control (Kan & Fabrigar, 2017). According to the theory, the determinants of behavioral intentions are a combination of three factors: attitudes toward the behavior, subjective norms encompassing the behavior execution, and personal perceived behavioral control. A behavior is an individual's overt action or set of actions that they perform and is confined to the researcher's theoretical and/or applied objectives and can be conceptualized in terms of an individual's action, target, and context, and the time it is performed (Kan & Fabrigar, 2017). The behavior's attitude is an individual's evaluation of performing the behavior as defined by these four components and assessed by measuring one's behavioral beliefs. Subjective norms are determined by normative beliefs and the motivation to conform with specific referents (Kan & Fabrigar, 2017). The normative beliefs are an individual's perception of with regards to expectations of the people important to them. Perceived behavioral control refers to people's perception of the ease or difficulty of directly performing the behavior of interest and influencing intentions or behavior (Kan & Fabrigar, 2017).

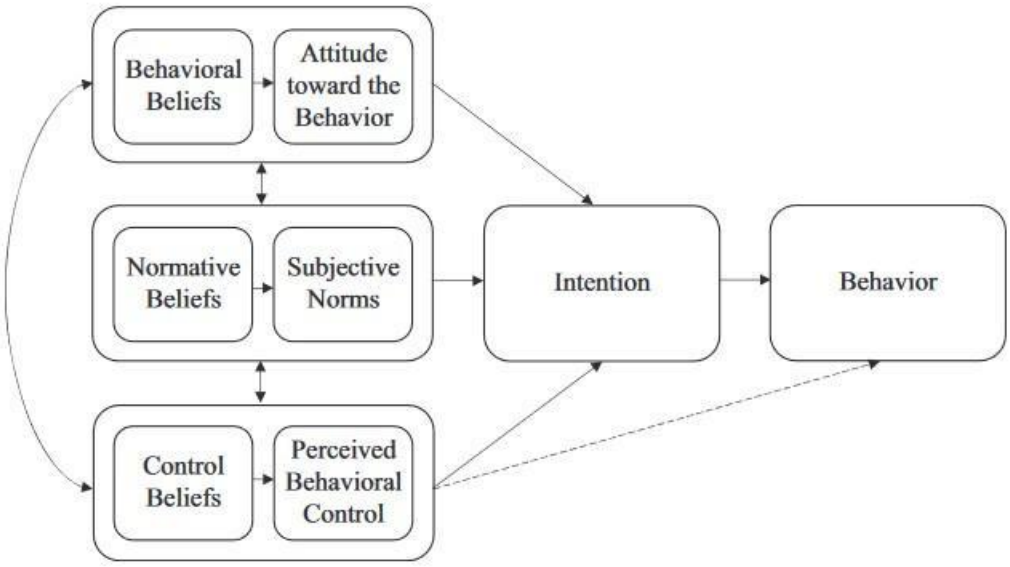

Figure 2:1 Theory of planned behavior
(Adapted from Kan & Fabrigar, 2017)

The theory of planned behavior suggests that sometimes people may intend to perform a given behavior but may lack complete control of their behaviors given the circumstantial current internal or external controls constraining the behavior under observation (Bhattacherjee, 2021). Internal controls constitute the person's ability to perform the intended behavior, while external controls are the availability of external resources needed to perform that behavior (Bhattacherjee, 2012).

Behaviors influence outcomes by partly contributing to the factors that lead to an outcome under observation. An outcome is a combination of a person's behavior and other external factors that positively and negatively influence the outcome.

Through this model, this study will investigate possible behaviors that influence the pre-delinquency of credit cardholders using transaction-oriented activities such as credit-utilization ratio, persistence debt, and cash advances.

2.3 Cardholder characteristics that influence delinquency

The personal cardholder characteristics that influence delinquency can be categorized into demographic, socio-economic, and behavioral factors. According to Ming-Yen et al. (2013), age, income, occupation, and marital status were posited to be the socio-demographic factors that greatly influence credit cardholders' propensity to default. Older credit cardholders tend to be sound and conservative in their spending decisions while their young counterparts tend to accumulate debt as they perceive their debt to be a temporary state which they will repay in the future as they're still young (Ming-Yen et al., 2013). According to a study by Zainudin, Mahdzan, and Yeap (2019), most young adults are technology savvy and digitally sophisticated than their older counterparts due to great exposure to digital payments and online shopping. According to this study young adults were found to be highly indebted due to excessive credit card usage caused by the intensified tendency for instant gratification, compulsive shopping behavior, high spending power and high debt commitment. The study also noted that the young credit cardholders were likely to pay the minimum monthly payment required while 50% admitted to falling into delinquency. In their research, Kiarie, Nzuki, & Gichuhi (2013) found that there was a lower rate of default among married cardholders compared to cardholders who were single. On gender, a study by Ciunova-Shuleska (2012) noted that most credit cardholders were males with females having a high tendency to repay their debt on time. Income was also significant contributing factor of how much credit limit a cardholder can access and repay. According to Stavins (2020), high-income credit cardholders were observed to have significantly higher credit card balances but showed tendency to repay those balances each month. Lower-income credit cardholders who were less educated exhibited a pattern of carrying unpaid balances, falling into the convenient cardholders' bracket.

The behavioral characteristics are a factor of the repayment patterns that are influenced by the cardholder's attitude towards repayment of debt, issuer characteristics, policy, and regulatory issues, and the prevailing economic conditions. The repayment action is normative in the essence that the cardholder is expected to fulfill their contractual

agreement by paying up their dues. Singh, Rylander, & Mims (2018) conducted a study on college students' behavior and defined credit card repayments as partial, full, or minimum payments and payments made on time. Those credit cardholders who pay full amounts on time are known as convenient users or transactors. On the other hand, revolvers are those that make partial or minimal payments with the balance being carried forward to the next payment period and therefore incur interest fees and sometimes late payment fees. Revolvers can further be classified into non-defaulting, and defaulting revolvers, where the latter have defaulted severally in paying their credit. Convenient users usually show self-cure signs, while revolvers are the accounts of concern since the risk of default is high, and accounts are likely to default since the credit cardholder's willingness to repay their loan is arguably low. In their study, Barboza, Smith, and Boubacar (2017) found that missing payments is a leading and significant source of anxiety in all model specifications due to prior financial mismanagement that creates adverse spillover effects in the current billing cycle while partial payments indicated financial distress.

According to the study by Shapiro, & Burchell (2012), financial anxiety was found to be a contributing factor to the financial administration of their finances, which consequently influenced their attituded towards repayment of debt. Barboza, Smith, and Boubacar (2017) argued that anxiety is inevitable, and one can experience anxiety in multiple dimensions of their lives, which can be a motivating or negative stressor. The depressive nature of financial anxiety influences the wrong assessment of outcomes leading to poor choices, which are primarily contributed by easy access to credit cards, low levels of financial literacy, which create pervasive incentives to live beyond one's means, and financial difficulties (Barboza, Smith, and Boubacar, 2017). Some of these choices have a negative financial impact, including over-borrowing, accumulation of revolving debt, or failure to make full payments according to schedules.

Since credit cards offer cardholders the flexibility of how quickly they repay their credit balance, individuals can choose whether to repay in full at the end of each month or spread payments over a long period (Financial Conduct Authority, 2017). The minimum

payments made each month are based on the credit agreement with the issuer. This flexible nature of making minimum payments each month means that credit cardholders can carry a large balance for an extended period without significantly reducing the debt. Holding a credit card balance for an extended period can also signify that a customer may be trapped in a cycle of borrowing that they cannot afford to pay down (Financial Conduct Authority, 2017). In some circumstances, a credit cardholder making the monthly minimum repayment may have underlying financial difficulties obscured by the repayment pattern (Financial Conduct Authority, 2017).

2.4 Credit Card Consumption Patterns

A credit card is used to make payments, either in a card-present and card-not-present environments, which defines the cardholder's interactivity and the payment platform. A card-present transaction involves the cardholder interacting with their physical card to complete payment by swiping their card through a reader or when a payment device processes an EMV chip in the card. On the other hand, a card-not-present transaction involves the cardholder making payments in a payment processing platform such as an e-commerce site without physically presenting their credit card to complete the payment. In a card-not-present environment, a combination of the credit card physical information such as the primary account number (PAN), expiry date, and the card verification value (CVV) are used to verify and authenticate the transaction. Some issuers enforce a cardholder security protocol layer commonly referred to as 3D Secure to prevent fraudulent activities on one's account by sending a one-time PIN (OTP) to the cardholder's mobile device to complete the payment. In either of these two scenarios, credit card transactions fall into either the following types: purchase, pre-authorization, capture, void, reversal, and verify.

A purchase, commonly known as a sale, is the most common type of credit card transaction where a cardholder uses their card to pay for goods or services. In this transaction, the cardholder initiates the payment, either from a merchant's payment device or in a payment gateway where the card data and the transaction amount is

submitted to the acquiring bank network, which is the financial institution that is responsible for processing the transaction on behalf of the merchant. The acquiring bank then sends this transaction to the card network, which reroutes the transaction to the issuer's bank for approval. If approved, the merchant receives an approval code signifying the completion of a successful financial transaction. This transaction type reduces the cardholder credit limit, and the cardholder is obligated to settle the debt according to the credit agreement with the issuer.

A pre-authorization transaction is similar to purchase with the exemption that, the transaction amount is not debited from the credit card account, but rather, it is "reserved" for a given number of days, usually between 7 to 10 days, as a guarantee of the availability of funds to be claimed later using a capture transaction. If a pre-authorization is not captured during the allotted period, the funds are released back to the credit card account. This method is often used by gas stations, car rentals, and hotels where the merchant desires to ensure that the transaction amount is available before offering a service. A pre-authorization request doesn't affect the credit card limit as the charge is not incurred in this leg. A capture transaction is the second leg of the pre-authorization transaction and utilizes the approval code returned in the pre-authorization phase to complete this transaction. The capture amount can be less or up to the pre-authorized amount but cannot exceed the same. The net effect of both the pre-authorization and capture requests is similar to the purchase transaction, with the only difference being the time when the actual debit is done from the account. Therefore, when the capture transaction is actualized, the credit limit is affected in similar ways to purchase.

A refund performs a reversal of the purchase transaction, and the transaction amount is credited back to the credit card account less the transaction fees. It is similar to an internal funds transfer transaction as funds are moved from the merchant's pool account to the credit cardholder's account to zero out the purchase transaction. A refund increases the credit limit of the credit card account and reduces the outstanding payable debt. A verification transaction is a zero-amount transaction processed into a credit card account to ascertain for the first time the validity of the card in card-not-present scenarios where

a charge is to be incurred in the future. It's usually used by merchants offering subscription services that are payable using card-on-file processing architecture where the card information resides with the merchant for future payment of services. The number of verification transactions in a credit card account depicts the card's active use in making purchases in different e-commerce sites or the possibility of payments to be debited into the account sometime in the future.

Other transactions that can be analyzed that determine a change in the cardholder's behavior include cash advances from the credit card, first time over limits over a period, a sudden credit card activity after a period of inactivity, and recent direct debit cancellations. These transactions signify a shift of the average cardholder's behavior, which is likely to be caused by changes in their financial status or deliberate intentions resulting from lack of attention to their credit limits.

2.4.1 Card activity after a period of inactivity

Credit card inactivity occurs when the card is not used to perform any transactions over a given time horizon. Inactivity in this aspect falls within the defined period of inactivity by the issuer before the credit card account is closed. Inactivity is influenced by several factors including and not limited to the cardholder having several credit cards to their intention of using the credit card only for emergencies.

2.4.2 Bounced Payments

A payment transaction bounces when the transaction fails to complete successfully at the point of service due to the card's limits, and the transaction is said to be rejected. When a purchase or capture transaction is rejected, it usually reflects that the cardholder has exhausted their credit card limit and that an increase of credit limit at the time of performing a transaction has been denied for an account with no protected overdraft facility. Rejected transactions are also an attribute of overspending, which is a measure of whether a cardholder knowingly transacts and does not have sufficient balance to fund the transaction.

2.4.3 Over limit for the first time in each time horizon

Over limits refers to when a credit card account surpasses its credit limit with a transaction. When the transaction amount exceeds the credit limit, the issuer may choose to decline the transaction, leading to a bounced payment or charge over-limit fees.

2.4.4 Cash Advances

A cash advance occurs when a cardholder withdraws money from an ATM or performs unprotected overdraft payments for issuers with credit overdraft facilities. Cash advances usually attract higher interests than normal purchases, have no grace period, and attract a direct cash advance fee. Cash advances usually follow the need to have cash that could have otherwise been withdrawn from one's current or savings account.

2.4.5 Change in the types of products or services purchased

Each financial message in the network is transmitted with the merchant category code (MCC), which is an ISO 18245 4-digit code that distinguishes the industry in which a merchant operates regarding the goods or services they provide, and this makes it easy to categorize the products or services a cardholder spends on. The type of merchant, the amount spent at each, and how often a cardholder shops with the card are key attributes that can reflect a significant change in the financial status of a cardholder when observed over time. When a cardholder who uses their credit card for purchasing non-essential services suddenly begins to buy essential goods or services such as groceries, this sudden change can be observed and analyzed as a factor that the cardholder's financial status has changed.

2.4.6 Credit utilization ratio

The credit utilization ratio is the credit card balance compared to the credit limit and determines the card's balance level. As the cardholder makes more purchases, the credit utilization ratio goes up, and so does the default risk. A cardholder with a low credit utilization ratio has a lower chance of missing their payments than one with a higher ratio when observed on the same time horizon. Ambrose et al. (2006) observed that credit utilization increases during periods of economic distress and noticed a trend in that initial credit utilization was lower for borrowers who expected future financial credit deterioration, while their counterparts utilized a more significant percentage of the total credit available.

2.4.7 Frequency of direct debit cancellation

Direct debits are payment arrangements that cardholders enter with a bank to facilitate the transfer of money from their other accounts to their credit card account to settle their outstanding debts on agreed debts. Credit cardholders are at liberty to cancel direct debits per the arrangements they have with their banks. Cancellation of direct debit before settling the credit card debt at the end of the billing cycle informs the intention of preserving access to cash on the checking account due to factors such as hoarding cash due to an uncertain economic difficulty. Simultaneously, it shows a low probability of settling card debt, mostly if this was their main checking arrangement.

2.5 Empirical Review

2.5.1 Single Classifier Algorithms

Single classifiers can be utilized to solve a credit scoring problem, which is a binary classification of two classes – "good" and "bad" customers (Tsai & Hung, 2014). Single classifiers can be categorized into two: statistical techniques and machine learning methods. The most used statistical techniques in credit scoring include linear discriminant analysis, multiple discriminant analysis and logistic regression among

others. Machine learning techniques include artificial neural networks (ANN), K-nearest neighbor, and support vector machines.

2.5.1.1 Discriminant Analysis

Fisher developed discriminant analysis in 1946 as a classical model for separating two groups using a linear combination of variables by deriving a linear combination of explanatory variables(ratios) that provides the maximum distance between the means of the two-subsets (Khemais, Nesrine, & Mohamed, 2016). In scoring, discriminant analysis can is used to separate credit cardholders into two subsets, those who are likely to miss payments, and those who will self-cure over the same period. This method allows the characteristics of those credit cardholders who are likely to miss payments to be determined and distinguished from those who will self-cure, and from these, a new cardholder pre-delinquency score can be determined. A discriminant analysis function for this model can therefore be presented as follows:

$$S_i = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

Equation 2:1 Discriminant analysis function

where a_1, a_2, …, a_n are the discriminant coefficients, X_1, X_2, …, X_n are the discriminant variables and S_i is the variable defining each of the subsets. When the S score is high, the cardholder is likely to miss payments, and when the S score is low, the cardholder is likely to self-cure.

While discriminant analysis is easy to implement and straightforward in generating results, one of its weaknesses is the assumption that the variables used normally distributed and independent, which is restrictive, is real-life practice since the credit data is categorical in nature. It also suffers from the fact that the subsets' covariances under observation might not be equal (Mittal, Gupta, & Jain, 2011).

2.5.1.2 Logistic Regression

Logistic regression is a linear model that estimates the probability of discrete outcomes based on several predictor variables. The most common logistic regression models binary outcomes bounded by the range 0 or 1; yes/no, success/failure, and will occur/won't occur. Therefore, logistic regression can predict a customer's probability to default and identify the variables related to this behavior. The mathematic representation of logistic regression is as follows:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_k X_{ji}$$

Equation 2:2 Logistic Regression Equation

(Adapted from de Paula et al., 2019)

where p_i is the probability of the customer i being good, k is the number of independent variables in the model, X_{ji} is the value of the independent variable j for customer i and β_k are parameters, i = 1, ..., k.

According to Regis and Artes (2015), logistic regression is used to estimate the probability of a customer transitioning from one state A, non-defaulting, to state B, defaulting over a certain period, and other kinds of transitions are not considered. However, a customer can possess other states other than these two, including non-defaulting without revolving credit, non-defaulting with revolving credit use, in delay, voluntary cancellation, and default (Regis & Artes, 2015).

2.5.1.3 K-Nearest Neighbor

The k-nearest neighbor classifier is a lazy, non-parametric, and instance-based learning algorithm used for classification and regression. Its non-parametric characteristic enables this technique not to make any assumptions in the underlying data, which is a useful attribute in real-life practice. According to Khemais, Nesrine, & Mohamed (2016), this technique assesses the similarities between the pattern identified in the training set and

the input pattern. The non-parametric nature of this method gives it an advantage of modeling irregularities in the risk function over the feature space (Khemais, Nesrine, & Mohamed, 2016).

2.5.1.4 Artificial Neural Networks (ANN)

ANN is a computational model originally inspired by how the human brain processes information from cognitive and computer science research developments. ANN mimics the human brain process by allowing for concurrent complex processing of inputs to achieve an output (Ala'raj, Abbod, & Radi, 2018). An ANN model consists of a network of mutually connected artificial neurons that transmit information to one another. The neurons are represented by some state (0 or 1), and each node may also have some weight assigned to them that defines its strength or importance in the model (Kaur & Kumari, 2020). The neurons are organized in layers, with an input layer representing a given input data vector, the hidden middle layers, and finally, the output layer which provides the desired output of classification. The main advantage of ANN is that the model does not take any prior assumptions about data distribution before learning, which significantly promotes the usability of ANNs in various applications.

2.5.1.5 Support Vector Machines (SVMs)

Support vector machine (SVM) is a machine learning technique used in both classification and regression. In an SVM linear classifier model, a data point is viewed as a p-dimensional vector separated by a maximum of p-1 hyperplanes (Kaur & Kumari, 2020). As many hyperplanes may be used to classify the data, the best hyperplane with the maximum margin between the two classes it separates is selected (Kaur & Kumari, 2020). This hyperplane is called the maximum margin hyperplane. In credit scoring, it is used to find an optimal hyperplane that categorizes the training input data into two classes – good or bad (Ala'raj, Abbod, & Radi, 2018).

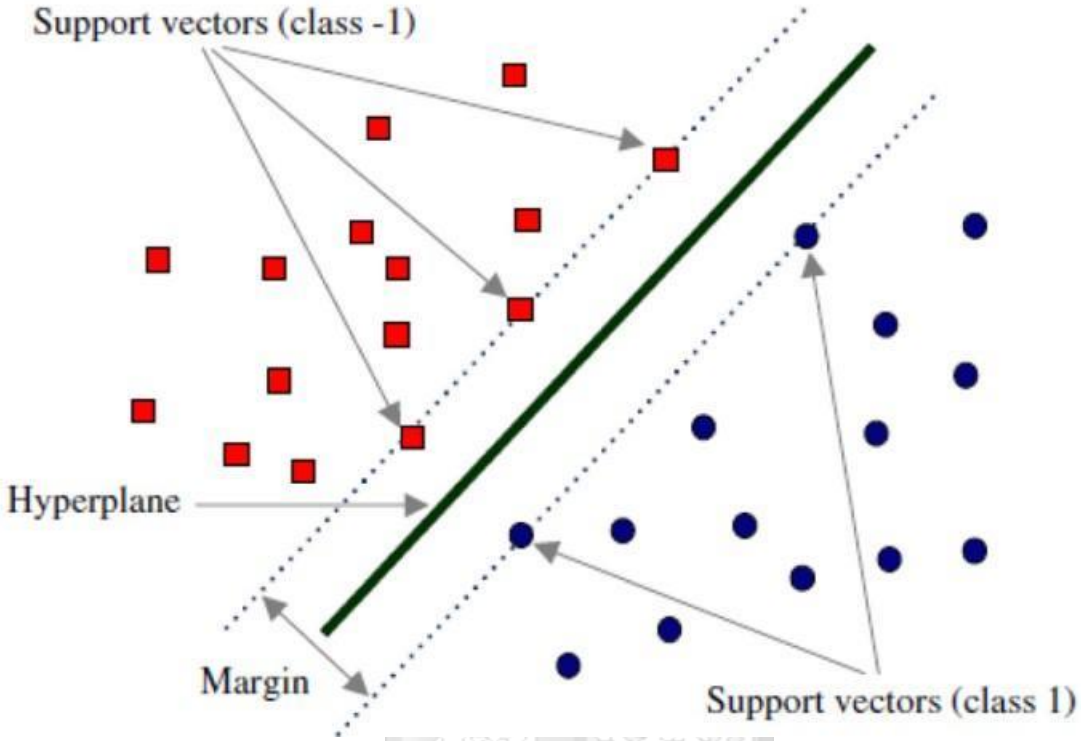

Figure 2:2 Support Vector Machines

(Ala'raj, Abbod, & Radi, 2018)

The dashed lines are called margins from the above figure, and the training data that lie on the margins are called support vectors. SVM works best when there is a clear margin of separation between classes, and therefore, is more effective in high dimensional spaces. However, it is not suitable where the dataset is considerably large or when the dataset has more noise.

2.5.2 Ensemble Classifier Algorithms

Ensemble learning is one of the techniques used to optimize and improve the performance of individual machine learning classifiers. An ensemble classifier consists of multiple base models that are trained, and their predictions combined to achieve higher predictive performance and reduce generalization error. The number of base models is

however kept small due to the computational training expense and diminishing returns of the final model performance as more models are used. Different ensemble techniques are chosen depending on the complexity of the problem at hand: varying the training data, varying the base models, or varying the combinations in generating the final ensemble output.

Machine learning model errors are defined by two properties: bias and variance. Bias is a measure of how close the model can capture the mapping function between inputs and outputs whereas variance refers to the amount by which the model changes when different training data is used in fitting (Zhanga & Ma, 2012). This introduces a trade-off with respect to the performance of a model – increasing the variance reduces the bias and increasing the bias reduces the variance of the model. Ensembles, therefore, achieve a better predictive performance in a prediction model than a single classifier by reducing the variance of the prediction error (Zhanga & Ma, 2012).

According to Tsai & Hang (2014), ensemble approaches can be classified into 3 categories: bagging, boosting, and stacking.

2.5.2.1 Bagging

Bagging classifier is an ensemble technique that Leo Breiman proposed in 1994 that handles classification and regression problems (Tsai & Hung, 2014). It is designed to improve the stability and accuracy of machine learning algorithms by combining classifications of randomly generated training sets to form a final prediction. Such techniques can be typically be used as a variance technique by randomization into its construction procedure and then creating an ensemble of out. Random forest is one of the popular implementations of the bagging technique which is used for both regression and classification based on a multitude of decision trees where each decision tree gives a classification (Hartmann, 2021). The algorithm uses a bootstrapping technique to train each tree in parallel on various subsets of the training dataset using different subsets of available features. To classify a new instance, each generated decision tree undergoes a voting process, and the final decision consists of the most popular class. Since each

decision tree in the random forest is unique, this reduces the algorithm's overall variance. This consequently enhances the predictive accuracy of the model and addresses the instability problem experienced in decision trees. Thus, they remedy the problem of individual decision trees which suffer from high variance and retain the benefit of a low-bias method making random forests to be appealing for both practitioners and researchers (Hartmann, 2021).

2.5.2.2 Boosting

Bosting involves learning from previous base model mistakes to make better predictions. It involves combing several weak base learners to form one strong learner, thereby improving the predictability of the models (Hartmann, 2021). It works by arranging weak learners in a sequence such that weak learners learn from the next learner in the sequence. The three widely applied boosting algorithms are Adaptive Boosting (AdaBoost), gradient boosting and Extreme Gradient Boosting (XGBoost). Adaboost combines multiple weak learners, which are decision trees with a single split known as decision stumps, to create a single strong learner that can be used for both regression and classification (Hartmann, 2021). Initially, all observations are weighted equally when creating the first decision stump. Observations that are incorrectly classified are then assigned more weights to correct the prediction error in successive iterations. In gradient boosting, predictors are added sequentially to an ensemble, with each set of preceding predictors correcting their successor which increases the accuracy of the model. New predictors are then fit to counter the residual errors in the previous predictor.

XGBoost is an implementation of gradient boosted trees that focus on the computational speed and performance of the target model. In XGBoost, the weak learners are added in parallel using a multithreaded pattern, which results in proper hardware utilization, greater speed, and efficiency. This algorithm is commonly used due to its ability to handle missing values and prevent overfitting, parallelization in constructing trees, out-of-core

optimization computing for large data sets, and cache optimization. XGBoost has been used successfully and across various Kaggle competitions (Hartmann, 2021).

2.5.2.3 Stacking

Stacking, or commonly known as stacking generalization involves integrating multiple classifiers with different classification algorithms (Tsai & Hung, 2014). In this method, a new model learns how to best combine the prediction outputs from other multiple existing models.

2.6 Conceptual Framework

The proposed model utilized a credit card dataset obtained from the UCI Machine Learning Repository that contained cardholder personal characteristics and transaction history that spanned over a period of 6 months. Exploratory data analysis was conducted to clean the data, identify patterns and relationships among variables. The dataset was first split into two – training and test – and then converted into weight of evidence (WoE) values into a scorecard. The training data was used to fit the scorecard using the XGBoost algorithm while the test dataset was reserved for validation purposes. The trained model would then be able to determine the pre-delinquency score in the defined billing cycle from those who were able to pay their credit card debt in full. The conceptual framework is illustrated with Figure 2.3 below:

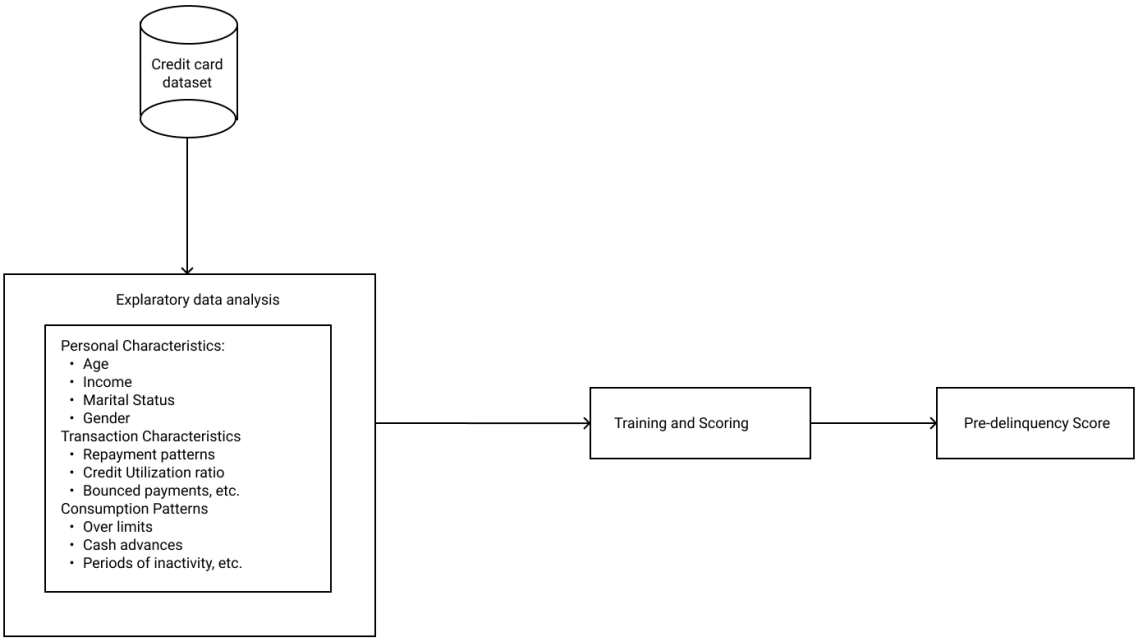

Figure 2:3 Conceptual Framework for the pre-delinquency prediction model

Chapter 3: Research Methodology

3.1 Introduction

This study aimed to develop a model for predicting credit cardholders' pre-delinquency using behavioral characteristics for a given time horizon. This chapter outlines the research methodology that was used to conduct the study.

3.2 Research Design

The study took an exploratory research approach to establish the relationship between the variables under observation and how they relate to the cardholder's propensity to repay their card debt. It also enabled assigning weights to the variables depending on their influence on their outcome in determining a credit cardholder's pre-delinquency level.

3.3 System Development Methodology

The research used an agile methodology, which is both iterative and incremental. This methodology takes a fail-fast approach to fast-prototyping that allows the model to be tested and validated as the system features are continually defined. This was suitable for a machine learning project as they involve a high level of uncertainty and allows the understanding of the business problem to be refined with each iteration through feedforward, feature re-engineering, and modeling.

3.4 System Analysis

The system is intended to be used by banks, and other financial institutions that issue credit cards and the system's main users will be the credit and collection officers. The system's functional requirements included allowing users to import data using CSV format, validating, and analyzing imported data, predicting the score of credit cardholders, and presenting the results in a user-friendly format. The non-functional

requirements included secure access to the system through proper authorization and authentication of users, ease of scale, and acceptable performance speeds. The system has a user interface from which all user intended actions will be submitted and results obtained.

3.5 System Design

Various tools were used to capture system interactions between the different components of the system. Use case diagrams were used to describe the different actors in the system and the main use cases that were involved to achieve their system needs. For each use case defined, sequence diagrams were used to explore and show the order of interactivity among the system's various components. An entity-relationship diagram defined the relationships of the different entity types that are to be modeled in the system's design.

3.6 System Implementation

The study used publicly available datasets containing relevant credit card data information to build the model since obtaining financial data from banks, and financial institutions are difficult due to its sensitive nature and the legal privacy obligation tied to them. The Python language was used for data manipulation, data analysis, and running the machine learning algorithms. MySQL community edition, an open-source database, was used for persistent storage and provides fault-tolerant, scalable, and high availability capabilities required for processing financial data. The ReactJs Framework was used to develop the front-end web interfaces that users will use to access the system due to its fast, scalable, and simple characteristics.

3.7 Target Population and Sampling

3.7.1 Population

The target population of this study was credit card cardholders. They constituted cardholders who have defaulted, with revolving balances, and those who have paid their

debt in full. The study also focused on credit cardholders with a repayment history over a period of 6 months. The source of this data was the UCI Machine Learning Repository which contains 30, 000 credit cardholders from a Taiwanese bank with transaction activities spanning from April 2005 to September 2005.

3.7.2 Sampling

The study focused on probability sampling to prevent bias and give each cardholder in the population an equal chance of being selected. This resulted in a sample with a true representation of the population. The study also adopted simple random sampling and from this, 80% of the collected records were used for training the model, 10% for testing, and 10% for validation purposes.

3.8 Data Collection

The study utilized a secondary dataset obtained from UCI Machine Learning Repository which contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card customers from a Taiwanese bank from April 2005 to September 2005. The dataset includes 30, 000 customer records and 24 attributes that will be significant in developing, testing, and validating the model. Given the scarcity of credit card datasets, this is the most recent publicly available data that suits the study.

3.9 Data Analysis Methods

The study performed an exploratory data analysis on the dataset to establish relationships and patterns in the demographic attributes, credit data, history of payment and bill statements over the course of 6 months. A histogram was used to visualize the proportion of credit cardholders who defaulted from those who didn't to determine the balanced nature of the dataset. Scatter plots were used to determine the distribution of payment amounts and bill amounts, as this was significant in understanding the consumption and repayment behavior of the credit cardholders. Scatter plots were also

used to visualize the distribution of age, gender, limiting balance and tendency to default to understand the influence of these variables and default. Stacked bar graphs were used to visualize the distribution of age and repayment amounts with the tendency to default.

3.10 Research Quality

3.10.1 Reliability

Reliability is the degree of consistency to which the model will yield the same results for the given set of inputs. This study will measure performance by evaluating the accuracy of the model, the precision, and recall ratios. Accuracy gives the ratio of the correctly classified observations to the total observations, which is given by the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 3:1 Model Accuracy

Where TP (True Positives), which are the correctly predicted positive observations; TN (True Negatives) is correctly predicted negative values; FP (False positives) which are wrongly predicted positive values; and FN (False Negatives) which are wrongly predicted negative values.

Precision is the ratio of all the correctly classified observations from the positive class among all the observations classified by the model as the positive class. It is expressed in the following formula:

$$Precision = \frac{TP}{TP + FP}$$

Equation 3:2 Precision Ratio

A recall is the ratio of the correctly classified observations from the positive class among all samples from the positive class in the actual data. It is expressed as follows:

$$Recall = \frac{TP}{TP + FN}$$

Equation 3:3 Recall Ratio

3.10.2 Validity

The research endeavored to maintain data integrity by using and presenting the training data as is and performing data preprocessing only to enable the model to work as intended. The data was not manipulated to produce baked results, and the results were presented as-is.

3.11 Ethical considerations

The study used publicly available secondary datasets and complied with the license agreements, privacy, and copyright dispute requirements under which the data was obtained. The study used both open-source and proprietary software with the latter being acquired through free educational licenses. The study proposal was also reviewed and approved by the Strathmore University Institutional Ethics Review Committee. The approval certificate for the study was received on 28th April 2021, and is attached in Appendix A.

Chapter 4: System Analysis, Design, and Architecture

4.1 Introduction

This chapter describes the general architecture and design of the system with an in-depth analysis of the proposed software solution. The study developed a web-based system that credit officers will interact with to generate real-time delinquency scores for their customers.

4.2 System Analysis

System analysis offered the researcher the capability to fulfill the software capabilities vis-a-vis the outlined objectives. The system's capabilities in this case is the software system that will enable credit officers to adequately score a credit cardholder in a given billing cycle and inform their decision of early interventions to avoid them falling into delinquency.

4.1.1 Requirements Gathering

4.1.1.1 Dataset Source

This study utilized a secondary dataset obtained from UCI Machine Learning Repository (accessible on https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients) due to the privacy and legal concerns associated with credit datasets from financial institutions that make it difficult to acquire such data. The dataset contained socio-economic and socio-demographic attributes such as income, age, gender, and marital status; credit card loan information such as the credit limit, and transaction history over a period of 6 months that included the monthly bills, the repayment amounts and the number of months in arrears, and most importantly, information on whether the credit

cardholders missed payments in the upcoming month. The default payment status is either 0 for credit cardholders who didn't miss payments, and 1, for those who missed. The cardholders' ability and willingness to pay is reflected in the pay amounts, and the number of months they are in arrears, and the ratios that can derived from these amounts. Therefore, this dataset presents the most up-to-date credit card information that contains relevant demographic and transaction history attributes that are significant to the study.

4.1.1.2 Data Dictionary for the Study

The dataset used contained 24 variables as described in the table below

Table 4:1 UCI Credit card dataset dictionary

SN.	Variable	Comments
1	ID	ID of credit cardholder
2	LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit
3	SEX	Gender (1=male, 2=female)
4	EDUCATION	1- Graduate School 2- University 3- High School 4- Others 5 & 6 Unknown
5	MARRIAGE	Marital status (1=married, 2=single, 3=others)
6	AGE	Age in Years
7	PAY_0	Repayment status in September 2005: Scale (the same applies to PAY_2 TO PAY_5) -1 = pay duly Other Positive Value – Payment delay for the value in months

		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_4	Repayment status in August 2005
9	PAY_3	Repayment status in July 2005
10	PAY_4	Repayment status in June 2005
11	PAY_5	Repayment status in May 2005
12	PAY_6	Repayment status in April 2005
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 in previous payment in August 2005(NT dollar)
21	PAY_AMT3	Amount in previous payment in July 2005(NT dollar)
22	PAY_AMT4	Amount in previous payment in June 2005(NT dollar)
23	PAY_AMT5	Amount in previous payment in May 2005(NT dollar)

24	default.payment.next.month	Default Status (0-No, 1-Yes)

4.1.1.3 Exploratory data analysis

Exploratory data analysis was performed to identify and define patterns and characteristics in the data without making any assumptions about what it might contain by using visual methods to highlight the narrative of the data. This provided an understanding of the most significant variables in the dataset and established a base for developing a parsimonious model. According to the dataset, 77.88% of the credit cardholders are not likely to miss payments in the next billing cycle, while 22.12% are likely to default. This indicates how imbalanced the dataset is and this must be accounted for when developing the model.

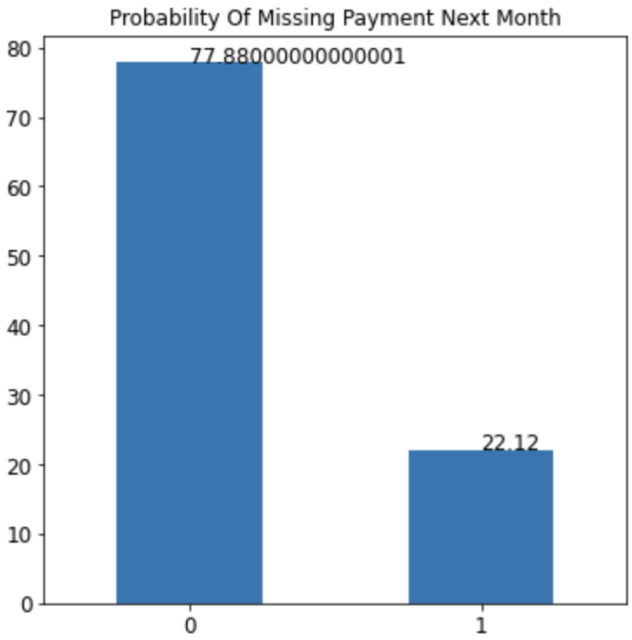

Figure 4:1 Probability of missing payment next month

The dataset consists of skewed data of limiting balance and age of credit cardholders as shown below. The plot in figure 4.4 shows there are more cardholders having a limiting balance between 0 and 200000, and with age between 20 and 40 years. The plot in figure 4.5 shows the highest number of customers fall between 21-30 and 31-40 age groups. Therefore, we can conclude that with the increasing age group, the number of customers that will default in the next payment month is decreasing. Thus, age is an important feature in predicting the probability of default.

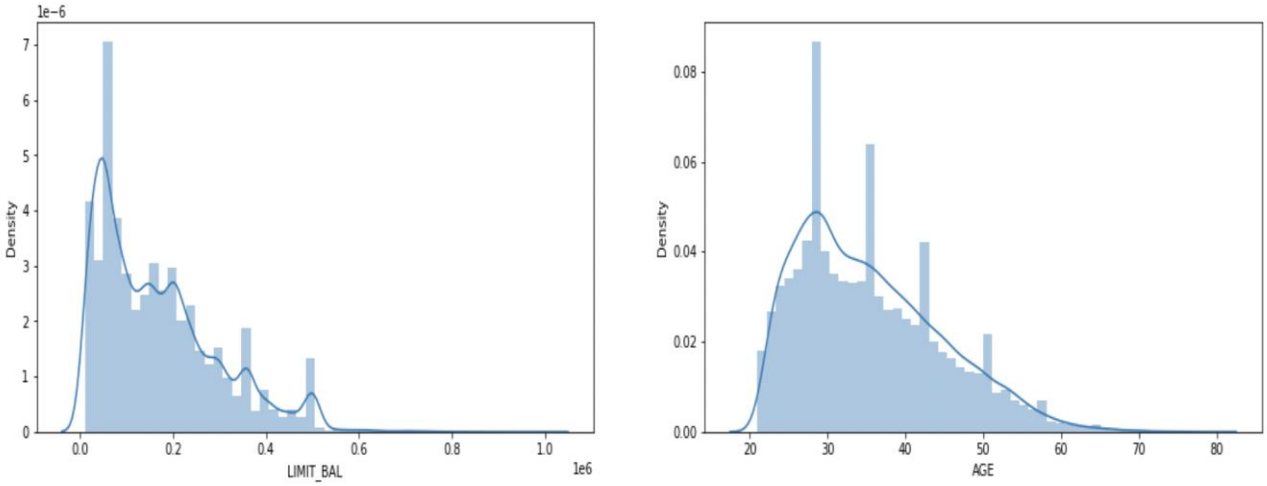

Figure 4:2 Distribution of age and limiting balance

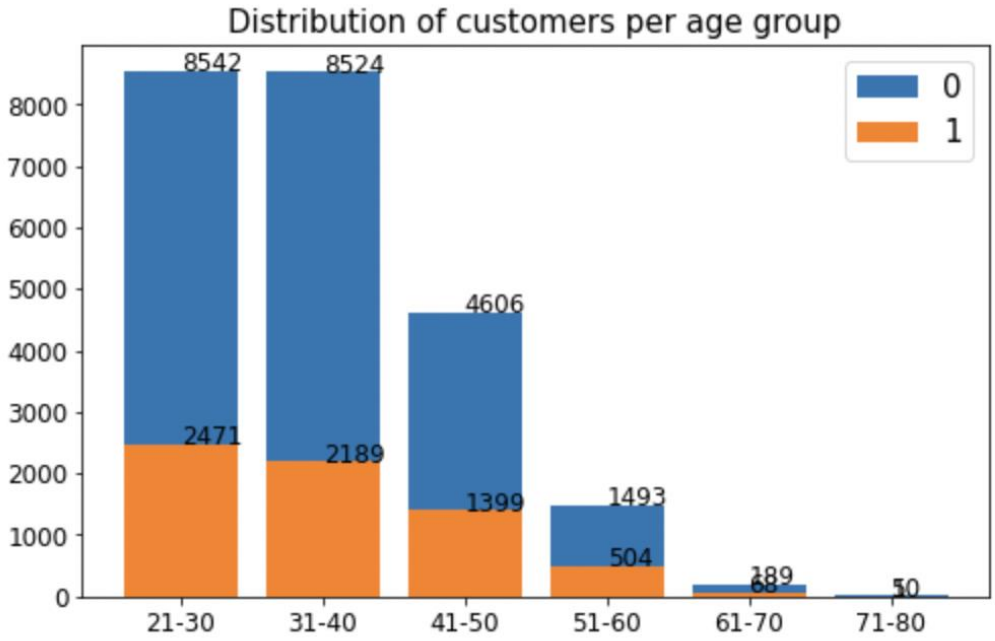

Figure 4:3 Distribution of customers per age group

The following plot shows the proportion of customers likely to default based on repayment history. We observe that, customers who paid on time had less probability of missing payments, while the probability of missing payments increased as the number of months in payment delay increased.

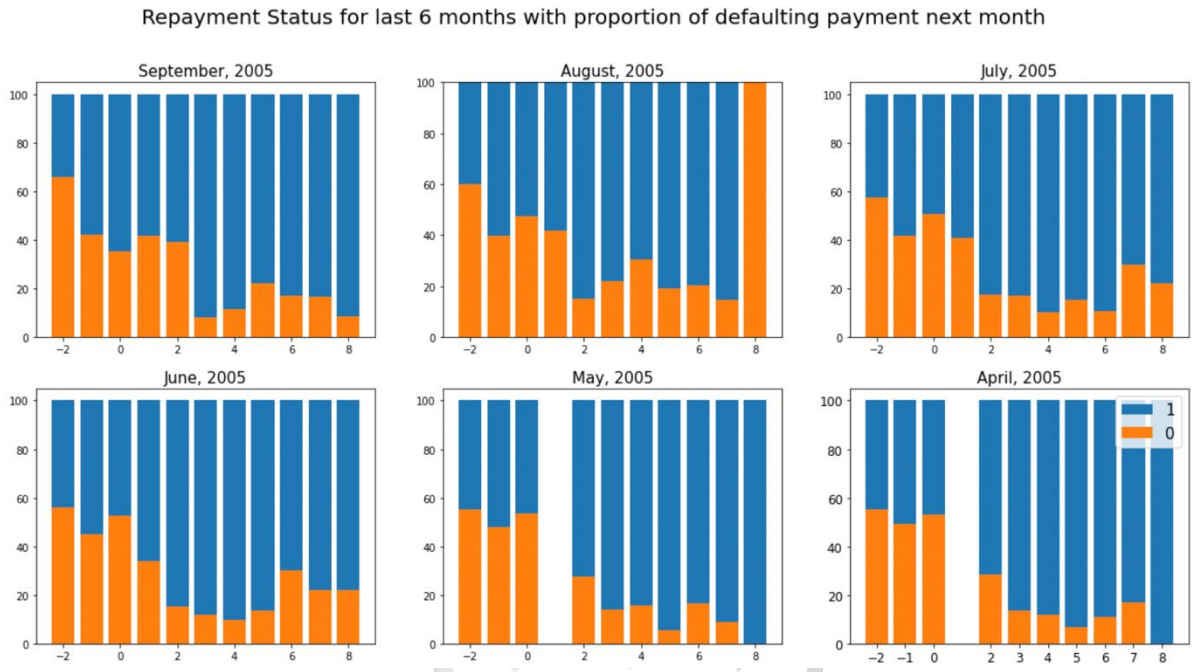

Figure 4:4 Repayment status for last 6 months vs probability of default

A plot of the distribution of age with marital status and the tendency to default showed that, customers within 30 and 50 age brackets and single customers of age 20-30 tend to default payment with the single customers having a higher probability to default.

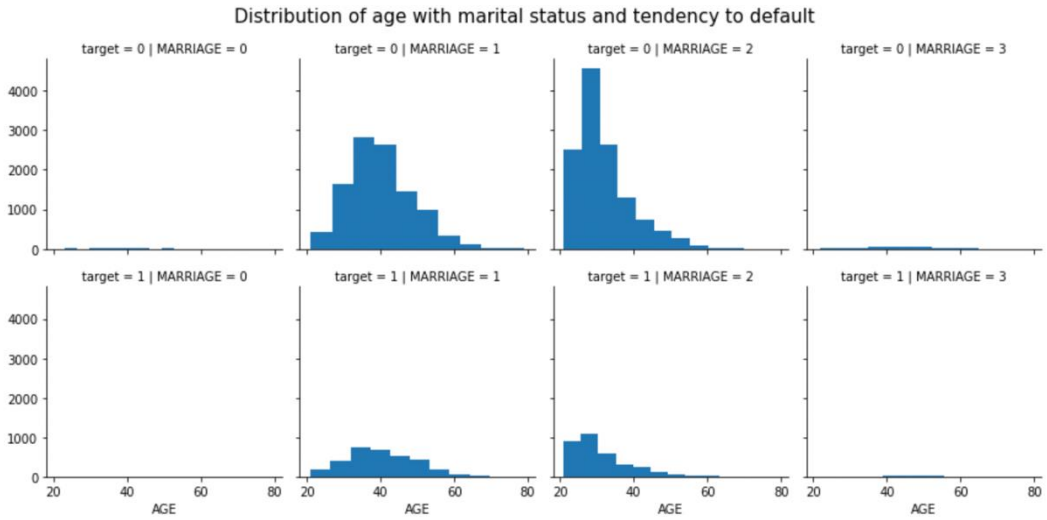

Figure 4:5 Distribution of age with marital status vs tendency to default

Females in the age group 20-30 had a very high tendency to default payment compared to males in all age brackets. In addition, the proportion of females who were not likely to default in the next billing cycle was also higher than the males.

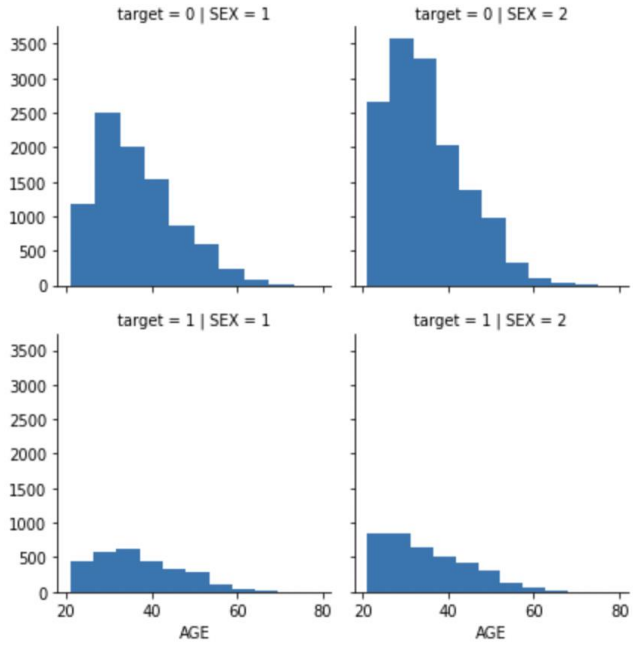

Figure 4:6 Distribution of sex and tendency to default

A scatter plot of the bill amounts against the payment amounts over the observation period revealed that there is a higher proportion of customers whose bill amount is high, but the payment done for the same billing month is very low. This is inferred by the concentration of datapoints along the y-axis and near to 0 on the x-axis:

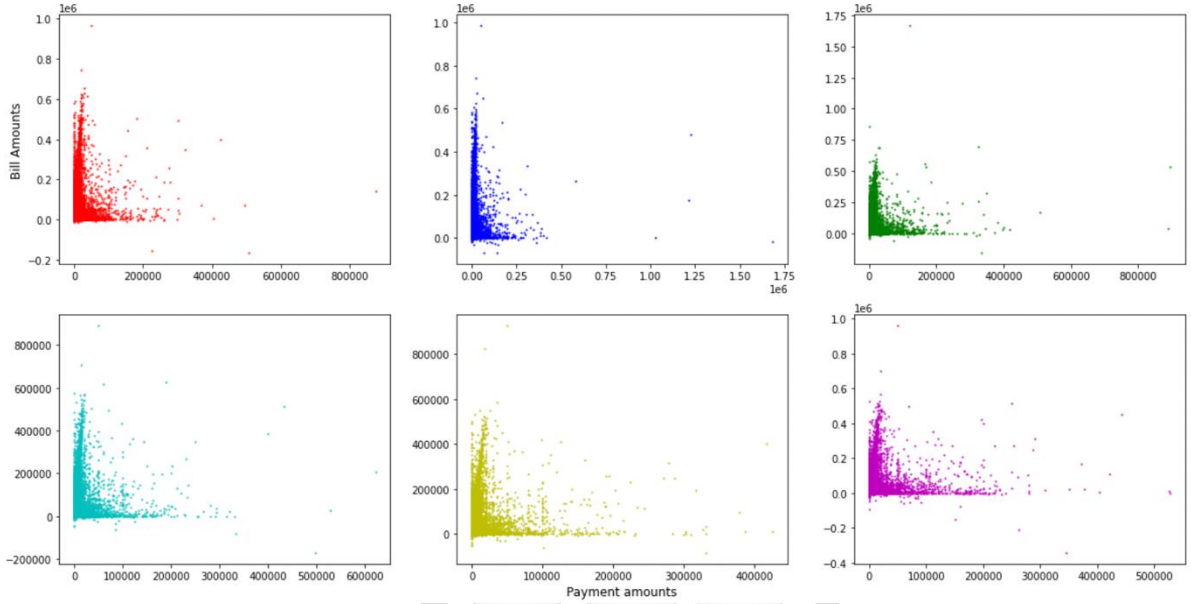

Figure 4:7 Distribution of bill amounts vs payment amounts

4.1.2 Functional Requirements

The system shall provide the following primary functions:

a. The system shall allow a credit officer to capture new customer details for scoring
b. The system shall perform relationships and patterns in the data to determine the characteristics that influence pre-delinquency
c. The system shall determine the pre-delinquency score of a credit card account

4.1.3 Non-Functional Requirements

Credit card data represents one of the most sensitive information that financial institutions must secure in day-to-day business operations. The system shall endeavor to fulfill the following:

a. System Security

The system shall follow secure coding standards, and high data security by complying with payments security requirements as defined in the Payment Card Industry Data Security Standard (PCI DSS) specification and other applicable security specifications.

b. Data Integrity

The system shall ensure accuracy, consistency, and completeness of data. Where applicable, the system shall track all events through audit logs, and enforce approval checks for the completion of system events.

c. System Reliability and Performance

The system shall provide high reliability and scalability and produce correct and consistent results in its life cycle.

d. System Performance

The system shall have high performance, with the in-built capability serving high traffic with low response time.

4.2 System Architecture

The system will consist of the frontend and backend components. The frontend responsibility will be to render the presentation logic to the users, who will sign in into the system and access the prediction form. The prediction form will allow the user to key in new credit cardholder customer data for prediction. These data will be sent to the backend using a REST protocol to the API component which will then sanitize and validate input and finally forward to the prediction engine. The prediction engine consists of an XGBoost classifier algorithm which will then generate prediction score, either 0 or 1, which shall be returned as a response by the API to the frontend component and presented to the user as feedback. This flow is captured in the architectural diagram

below:

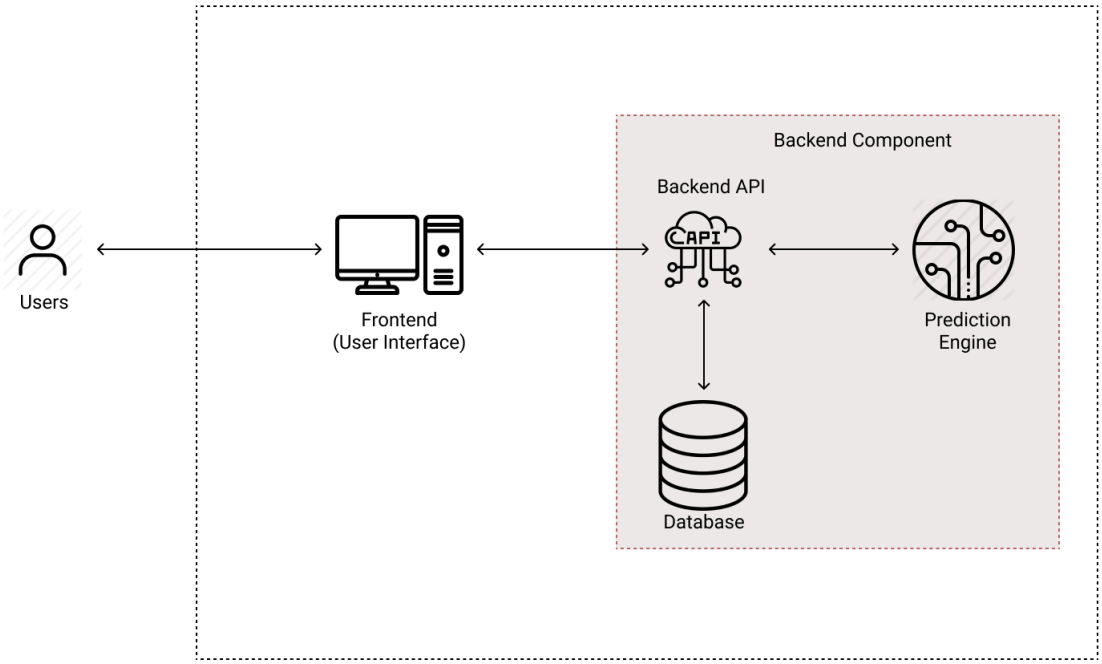

Figure 4:8 System Architecture

4.3 System Design

4.3.1 Use Case Diagrams

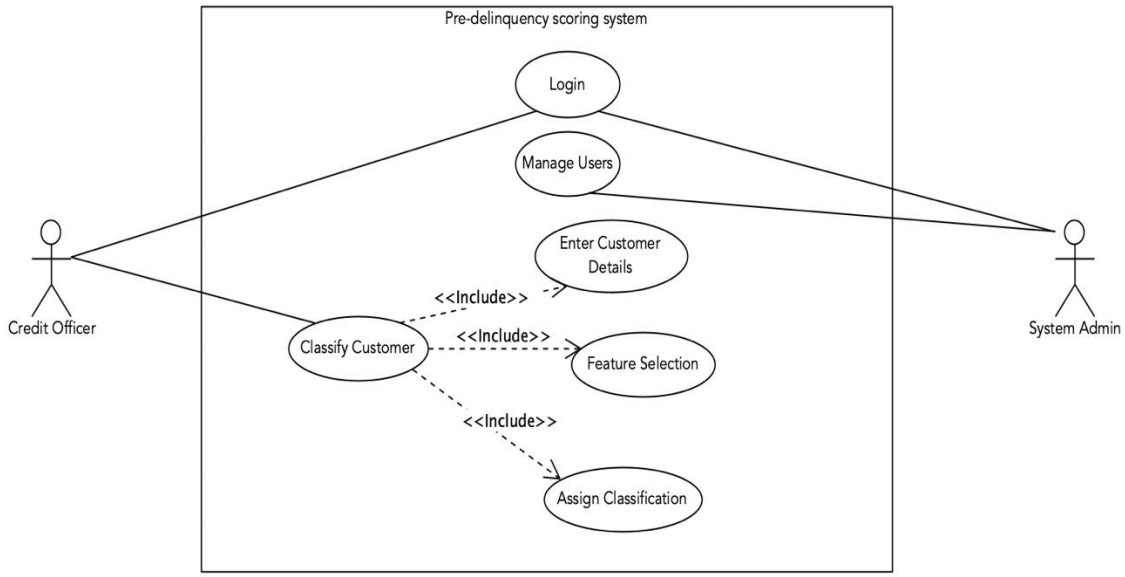

Figure 4:9 Use case diagram

Description of Use cases

This section provides detailed descriptions of the above use cases using the two-column fully dressed format:

Table 4:2 Use case UC1- Login

Use Case	Login
Primary Actors	Credit officer, System Admin
Post-conditions	User verification and access granted
Main Success Scenario	
Actor Intention	System Responsibility

1. User submits email and password combination	
	2. System authenticates users
Extensions	
At any time when the user forgets their password: a). A user will initiate a password recovery process by submitting their email address b). The system shall send an email notification with a link to set-up a new password c). The user shall submit their new password and attempt to login	

Table 4:3 Use case UC2 - Get Customer Classification

Use case:	Get Customer Classification
Primary Actors	Credit Officer
Pre-Conditions	User is successfully authenticated
Post-Conditions	Customer classification result
Main Success Scenario	
Actor actions	System Responsibility
1.User submits customer details	
	2. System performs feature selection
	3. System classifies customer with the given features
	4. Return customer classification

Table 4:4 Use case UC3 - Manage users

Use Case	Manage Users
Primary Actors	System Admin

Pre-Conditions	User is successfully authenticated
Post-conditions	User profiles are updated accordingly
Main Success Scenario	
Actor Actions	System Responsibility
1. Submits user profile information	
	2. System updates user profile information

4.3.2 Sequence Diagram

A sequence diagram highlights the order of operations from start to finish, and is depicted in the figure below:

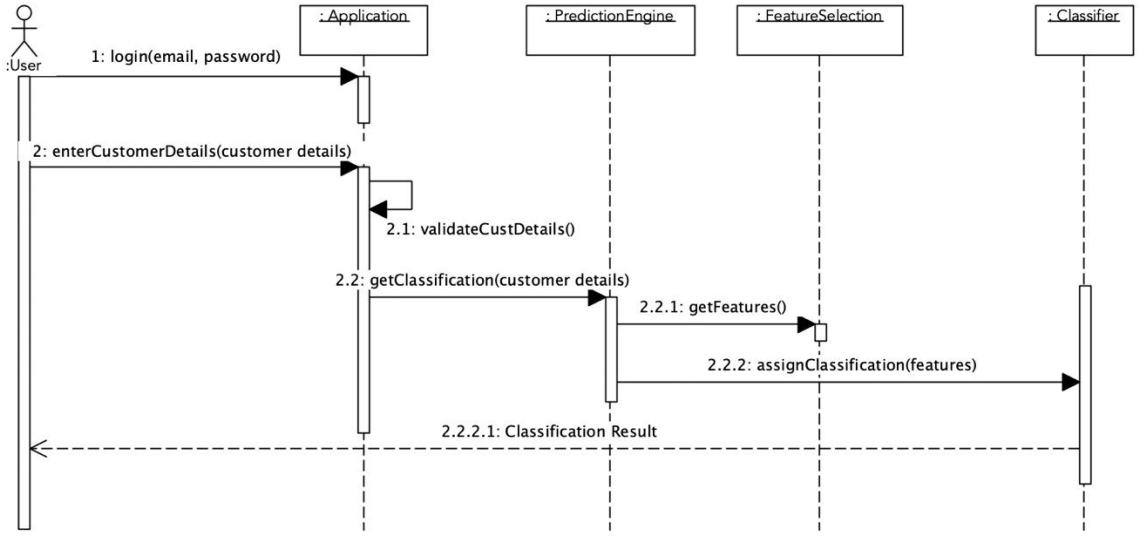

Figure 4:10 Sequence Diagram

4.3.3 ERD

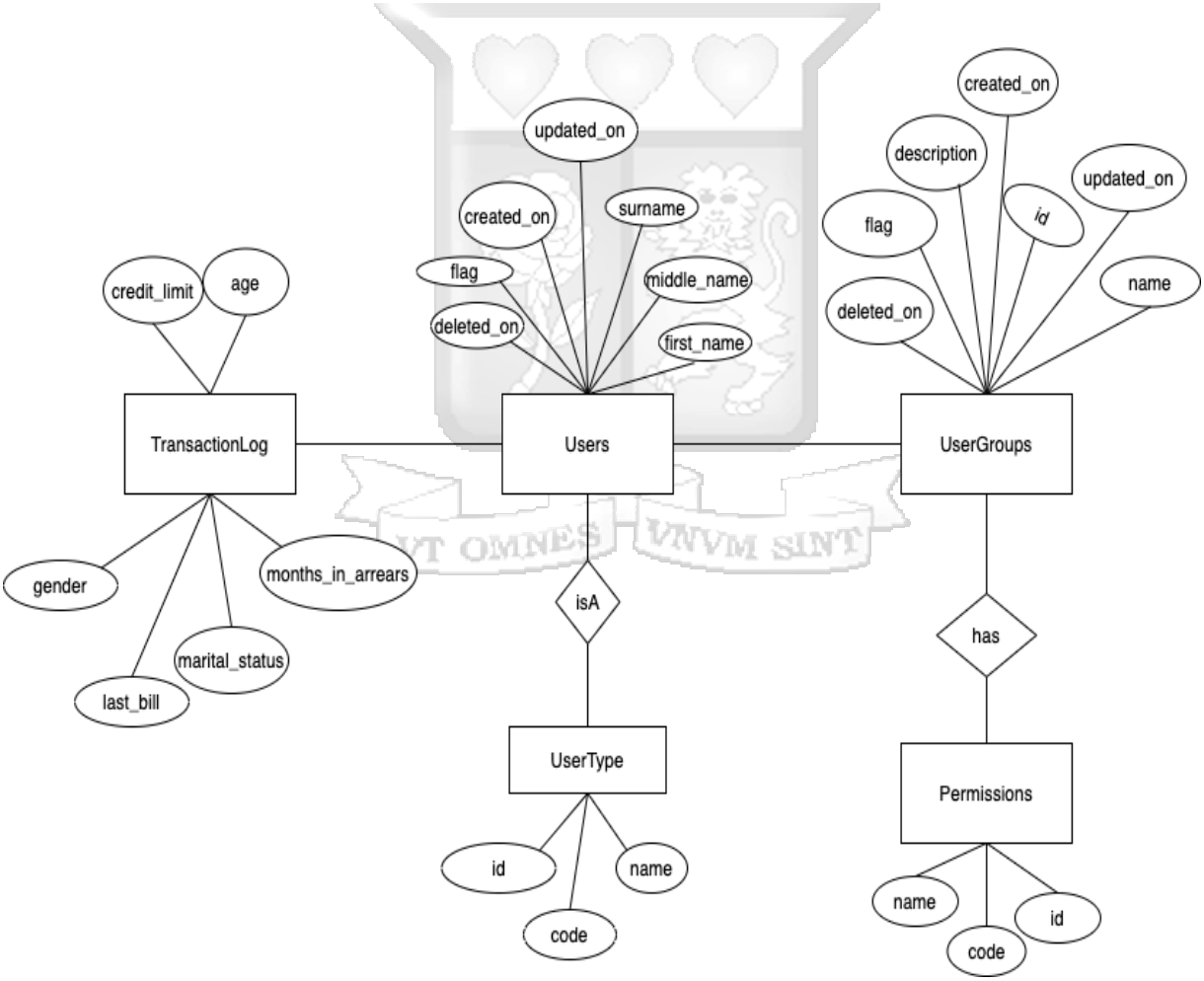

Figure 4:11 Entity relationship diagram (ERD)

4.3.4 Database Schema

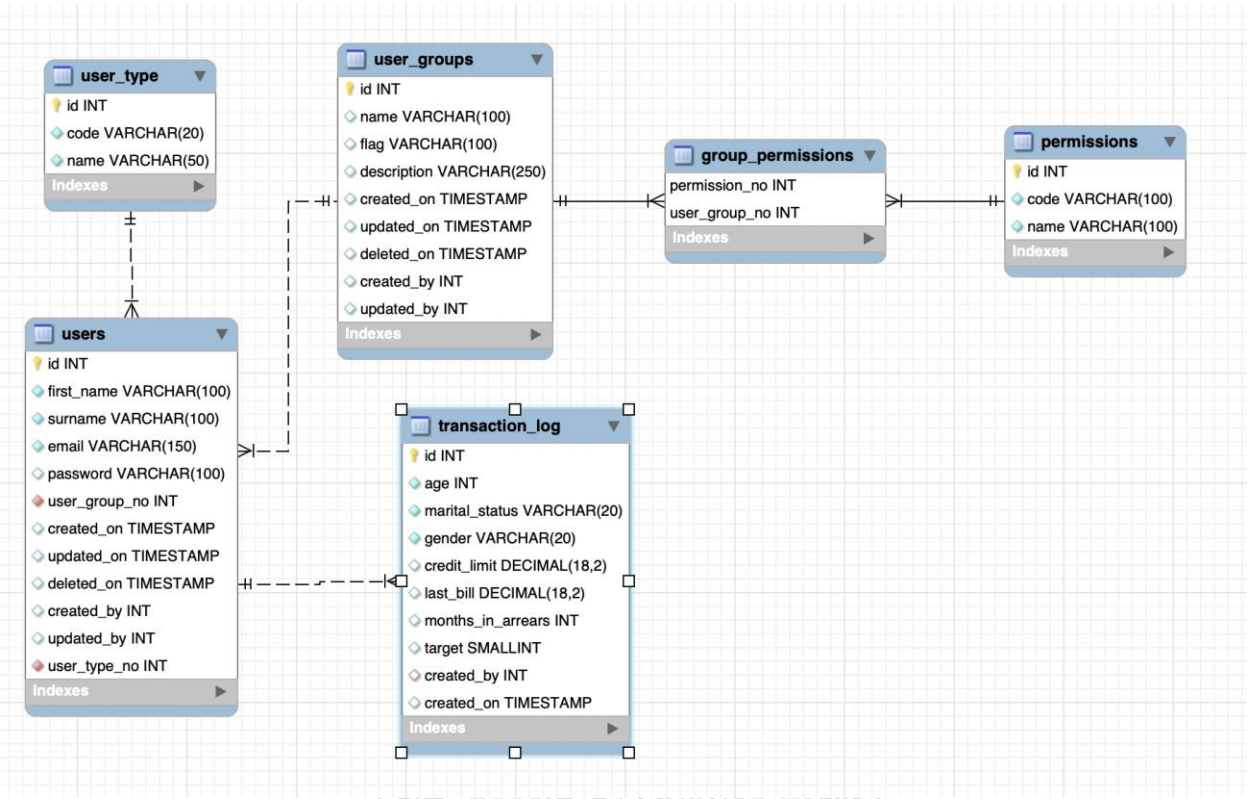

Figure 4:12 Database Schema

4.3.5 Class Diagram

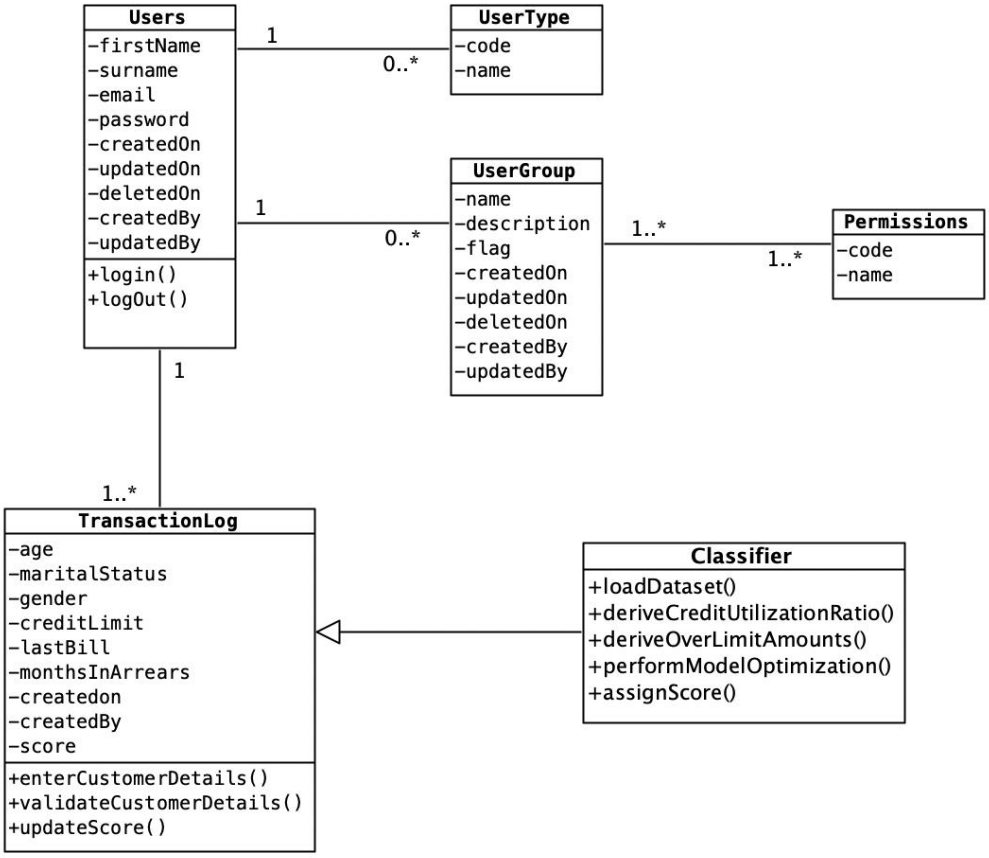

Figure 4:13 Class diagram

4.3.6 Wireframes of the system

A wireframe is a mock-up or a skeletal visual representation of the system for conceptualizing the system layout components. The following figure captures the login feature of the system:

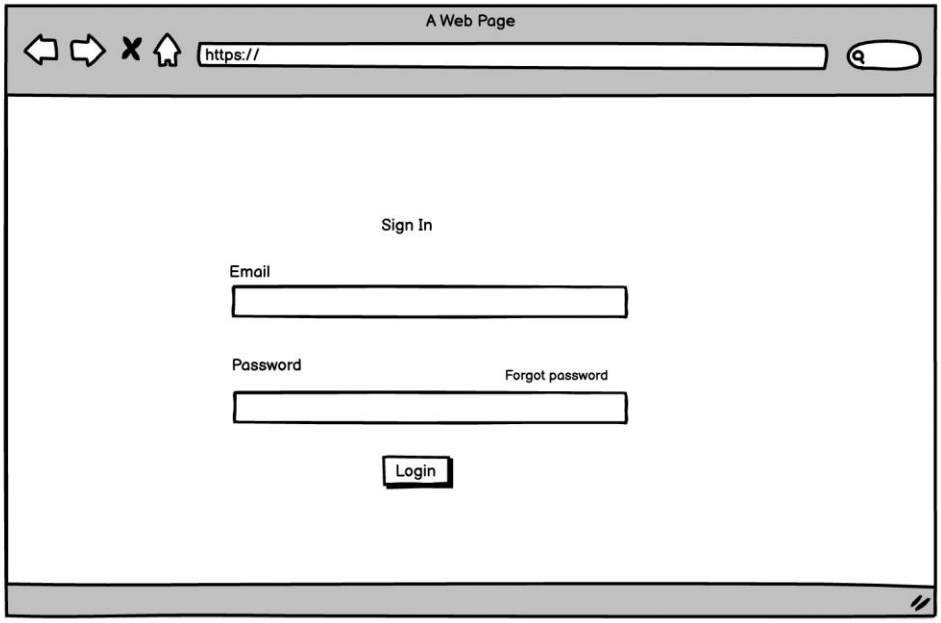

Figure 4:14 Login wireframe

Once a user has logged into the system, the system will redirect them to a screen that allows them to enter customer details and predict the likelihood of this customer missing payment in the current billing cycle. This is captured in the wireframe below:

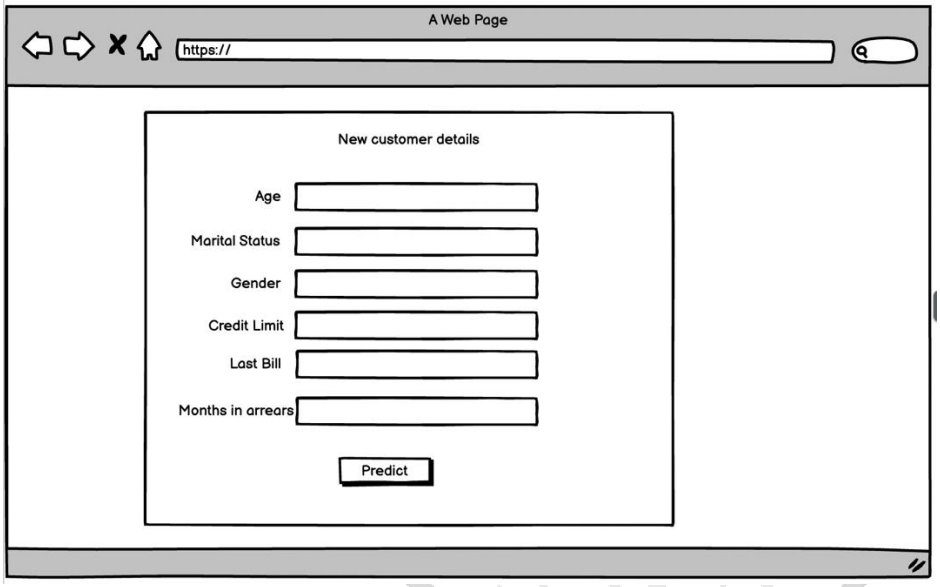

Figure 4:15 Prediction form wireframe

Chapter 5: System Implementation and Testing

5.1 Introduction

This chapter describes how the prediction system was developed, tested, and validated. The system consisted of a web solution with frontend, and backend modules. The frontend module provided users with graphical interfaces for interacting with the system. The backend module was an integration of a REST API interface serving the frontend requests and internal linkage to the prediction model. This chapter covers how the model was developed and how the system was integrated to form a web-based software product.

5.2 Model Development

The model was developed using the Python programming language and used the scikit-learn library for implementation. The scikit-learn is an open-source machine learning library that is extensive, accessible and provides high-level API interaction for most

machine learning methods and statistical modelling for classification, regression, clustering, and dimensionality reduction.

Data pre-processing was first performed as described in section 4.1.1.2 and included derivation of the credit utilization ratio and the overdraft amounts. The CSV dataset was first loaded into a panda's data frame as shown below:

```
def load_dataset():
    """ Load dataset from a CSV file"""

    file_path = 'uci_credit_card.csv'
    df = pd.read_csv(file_path)
    return df
```

Figure 5:1 Script for loading dataset

The credit utilization ratio is a measure of the total revolving credit already consumed to the total revolving credit available. It can also be expressed as the total debt divided by the credit limit in the case of a single credit card. This can be expressed using the following equation:

$$Credit\ Utilization\ Ratio = \frac{Total\ Debt}{Available\ Credit(Credit\ Limit)}$$

Equation 5:1 Credit utilization ratio

For this given dataset, the average credit utilization ratio was derived as follows:

$$Credit\ Utilization\ Ratio = \frac{BILL\_AMT1 + BILL\_AMT2 + BILL\_AMT3 + BILL\_AMT4 + BILL\_AMT5 + BILL\_AMT6}{Credit\ Limit * 6\ months}$$

Equation 5:2 Computation for credit utilization ratio

The credit utilization ratio in code was generated as follows:

```
def derive_credit_utilization_ratio(dataframe):
    """ Generate the average credit utilization ratio for the 6 month period """

    dataframe['credit_utilization'] = ((dataframe['BILL_AMT1'] + dataframe['BILL_AMT2'] + dataframe['BILL_AMT3'] + dataframe[
        'BILL_AMT4'] + dataframe['BILL_AMT5'] + dataframe['BILL_AMT5']) / (dataframe['LIMIT_BAL'] * 6))
    return dataframe
```

Figure 5:2 Script for credit utilization ratio function

The overdraft amounts for each of the 6 observation months were computed as follows:

```
def derive_overdraft_amounts(dataframe):
    """ An overdraft will be the value charged above the credit limit"""

    for counter in range(1, 7):
        derived_col_name = f"overdraft_amt{counter}"
        bill_col = f"BILL_AMT{counter}"
        dataframe[derived_col_name] = dataframe.apply(
            lambda row: (0 if row[bill_col] < row.LIMIT_BAL else row[bill_col] - row.LIMIT_BAL),
            axis=1)
    return dataframe
```

Figure 5:3 Script for overdraft function

5.2.1 Model Tuning

The XGBoost algorithm has several parameters that can influence the accuracy and the training speed of a model, and this study utilized the following parameters: the *n_estimators*, *early_stopping_rounds, learning_rate* and *n_jobs*. The value of *n_estimators* influence underfitting and overfitting of the model - a low *n_estimators* value causes underfitting which results to inaccurate predictions on both training and validation data, while a large value causes overfitting which results to accurate predictions on the training data but inaccurate predictions for any other set of data. The *early_stopping_rounds* define the number of model iterations that can be performed with significant improvements on

the validation score. A good training practice is to set a high value for *n_estimators* and then use the *early_stopping_rounds* to find the optimal time to stop iterating. The *learning_rate* is a multiplier that is used as a factor of prediction from each component model before summing up the predictions. A small *learning_rate* value is usually used since it yields more accurate models, though at the expense of longer training time. The *n_jobs* define the number of CPU machines that can be allocated to the model for parallelism to build models faster. A good practice is to set the value of *n_jobs* to be equal the number of processing cores in the host machine.

The Bayesian optimization parameter-tuning algorithm was used to obtain a best-optimized set of parameters due to its ability to constantly learn from previous optimizations and requires fewer samples to learn or derive the best values. A 10-fold validation was then performed for each specified set of parameters in each iteration. This is as shown in the figure below:

```
def xgb_evaluate(max_depth, gamma, n_estimators, colsample_bytree):
    """ Bayesian optimization function """

    def_params = {
        'eval_metric': ['rmse', 'auc'],
        'max_depth': int(max_depth),
        'n_estimators': int(n_estimators),
        'subsample': 0.8,
        'eta': 0.1,
        'gamma': gamma,
        'metrics': 'auc',
        'colsample_bytree': colsample_bytree}

    # 10-fold cross-fold validation with the specified parameters in 100 iterations
    cross_val_result = cv(def_params, dtrain, num_boost_round=100, nfold=10)

    # Since bayesian optimization only maximizes, return the negative RMSE
    return -1.0 * cross_val_result['test-rmse-mean'].iloc[-1]
```

Figure 5:4 Script for parameter model tuning

The best parameters were then extracted and used to fit the model as shown in the following figure:

```
xgb_bo = BayesianOptimization(xgb_evaluate, {'max_depth': (3, 7),
                                             'gamma': (0, 1),
                                             'n_estimators': (100, 120),
                                             'colsample_bytree': (0.3, 0.9)})

# Use the expected improvement acquisition function to handle negative numbers
# Optimally needs quite a few more initiation points and number of iterations
xgb_bo.maximize(init_points=3, n_iter=5, acq='ei')

# Extract the best parameters for the model
params = xgb_bo.max['params']

# Converting the max_depth and n_estimator values from float to int
params['max_depth'] = int(params['max_depth'])
params['n_estimators'] = int(params['n_estimators'])
params['booster'] = 'gblinear' #

# Generate the model
classifier = XGBClassifier(**params)
```

Figure 5:5 Script for Bayesian optimization

5.2.2 Scoring Method

The data was first transformed using the weight of evidence (WoE) method which attempts to find a monotonic relationship between the input features and the target variable by spitting each feature into bins and assigning a weight to each bin. This is based on the proportion of "bad" and "good" customers in each group level and measuring the strength of grouping for differentiating good and bad risk. Negative values usually indicate that a given bin group contains a higher proportion of "bad" customers than "good". The calculation for WoE can be expressed as follows:

$$WoE = ln\left\{\frac{Percentage\ of\ good\ customers}{Percentage\ of\ bad\ customers}\right\}$$

Equation 5:3 Weight of Evidence Equation

The *scorecardpy* library was used in implementing the scorecard as it was easy to use and efficient. The library has in-built functions for data partition, variable selection, weight of evidence binning, scorecard scaling and performance evaluation. The following script shows how the scorecard was implemented:

```
# filter variable via missing rate, iv, identical value rate
dt_s = sc.var_filter(df, y="target")

# Split the parsed data into train and test for the scorecard
train, test = sc.split_df(dt_s, 'target', ratio=0.8, seed=42).values()

# WoE binning
bins = sc.woebin(dt_s, y="target")

# Binning adjustment: using the interactive approach
breaks_adj = sc.woebin_adj(dt_s, "target", bins)
bins_adj = sc.woebin(dt_s, y="target", breaks_list=breaks_adj)

# Converting the train and test data into WoE value
train_woe = sc.woebin_ply(train, bins_adj)
test_woe = sc.woebin_ply(test, bins_adj)

y_train = train_woe.loc[:, 'target']
X_train = train_woe.loc[:, train_woe.columns != 'target']
y_test = test_woe.loc[:, 'target']
X_test = test_woe.loc[:, train_woe.columns != 'target']

classifier = classifier.fit(X_train, y_train)

# Generate a scorecard using the XGBoost classifier
card = sc.scorecard(bins_adj, classifier, X_train.columns)

# credit score
train_score = sc.scorecard_ply(train, card, print_step=0)
test_score = sc.scorecard_ply(test, card, print_step=0)

# psi
sc.perf_psi(
    score={'train': train_score, 'test': test_score},
    label={'train': y_train, 'test': y_test}
)
```

Figure 5:6 Scorecard model script

The model had a low population stability index (PSI) of 0.0005. The PSI measures changes in the score distribution against the development population and identifies shifts in population for the scorecard. A low PSI values infers an insignificant change, and that performance analysis of the scorecard is not required to be performed. The score distribution is as captured in the figure below

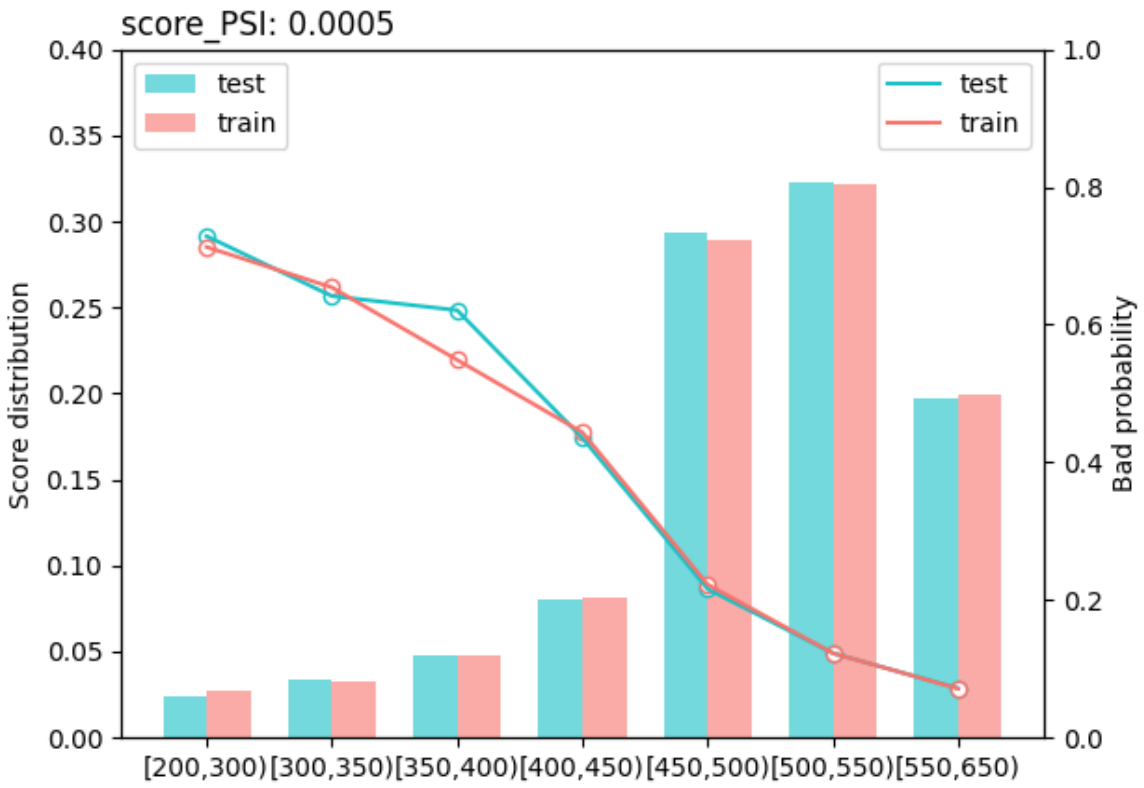

Figure 5:7 Score distribution

The predicted score ranged from 200-650. A low score, which in other words is a low predelinquency score, indicated that the customer has a lower probability of missing payment, while a high score signified that a customer had a high probability of missing

payment. The following score map was generated as a guiding principle for the various credit score brackets:

Table 5:1 Score risk mapping

Score	**Risk Category**
200-300	True Low Risk
300-400	Medium Risk
400-500	High Risk
500-650	Ultra-High Risk

5.3 System Implementation

The frontend module was implemented using React, which is an open-source JavaScript library for building user interfaces. It offers fast, scalable and simplicity capabilities when creating single page applications and the ability to re-use user interface components in a clean way. The frontend module allowed users to be authenticated using a login form, where they are required to submit valid authentication details, to which they're directed to the authorized resources as per their profile. System administrators have the capability to define system configurations and manage users, while credit officers have the capability to perform predictions based on customer details and receive feedback from the system.

For all these interactions, the frontend module makes asynchronous requests to the backend module through a REST API. The backend module was developed using the Python programming language and used the Flask framework which provides a minimalist approach in development, making it easy and fast to develop and deploy backend modules. A MySQL database server was used to store all the data relating to users, transaction logs – which are the predicted transaction events - and audit log trails.

The following figures capture the main screens that a credit officer must interact with to generate a prediction of a given customer:

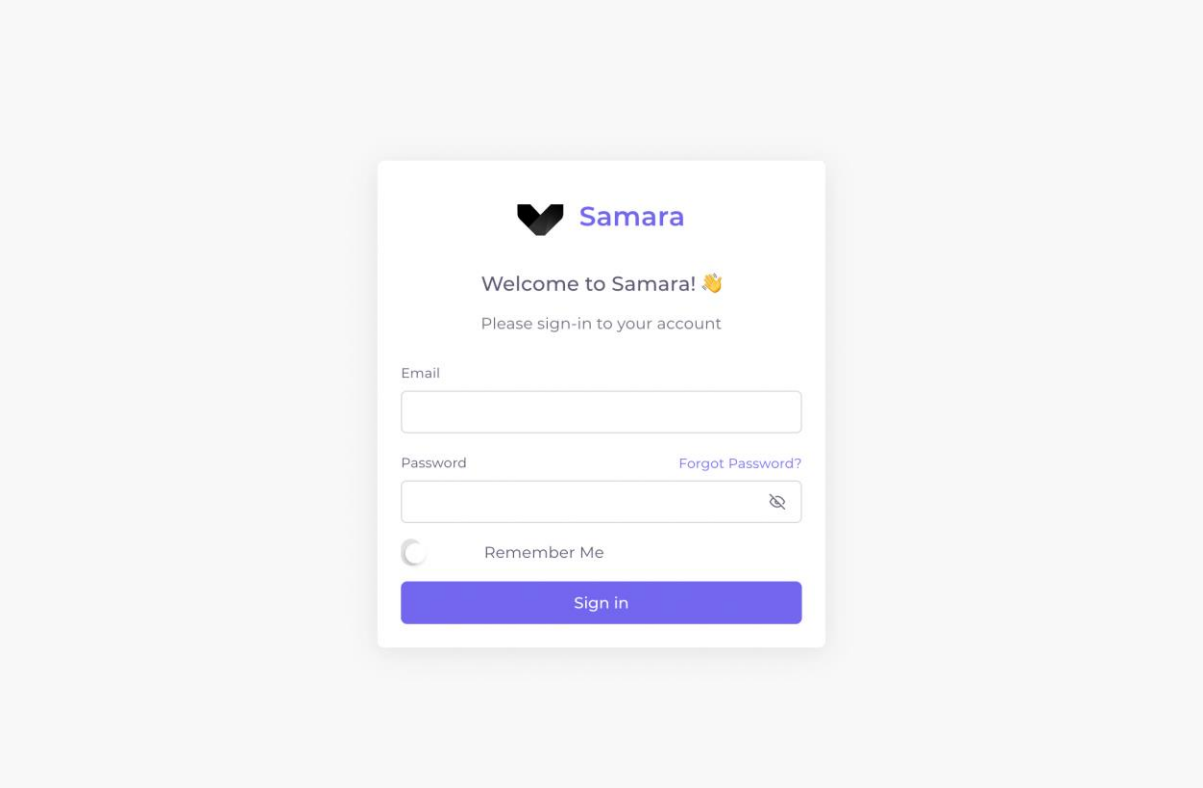

Figure 5:8 Login Screen

Prediction form

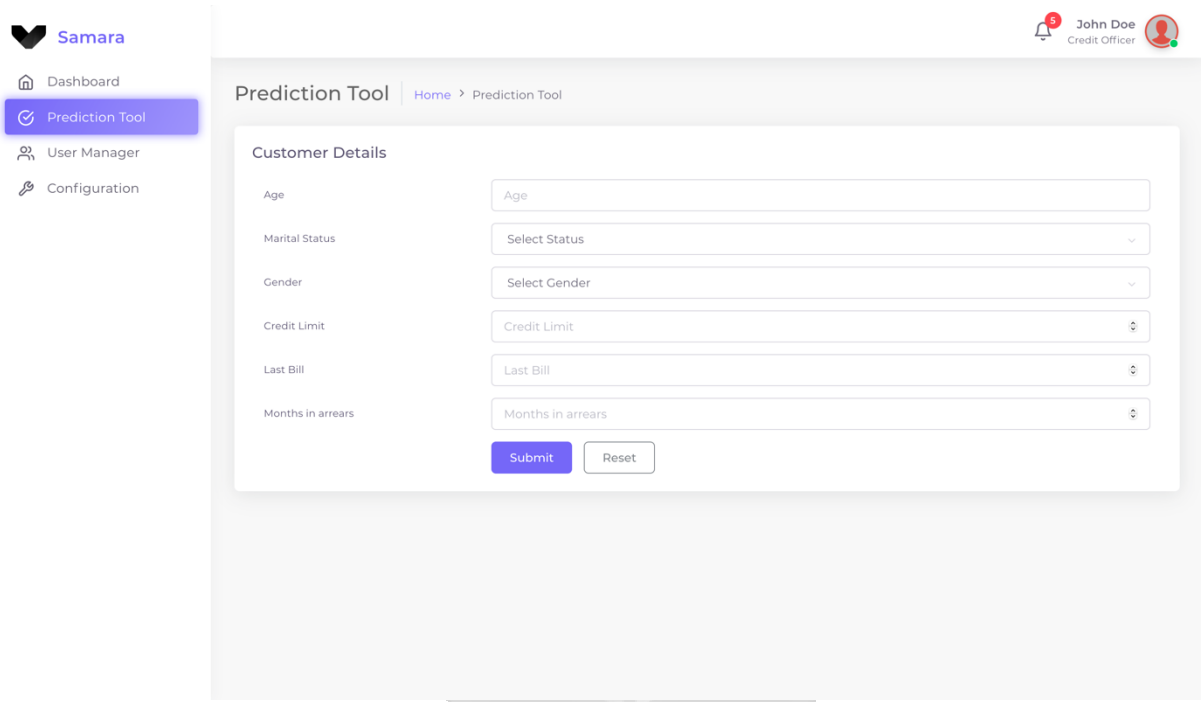

Figure 5:9 Prediction Form Screen

Prediction results

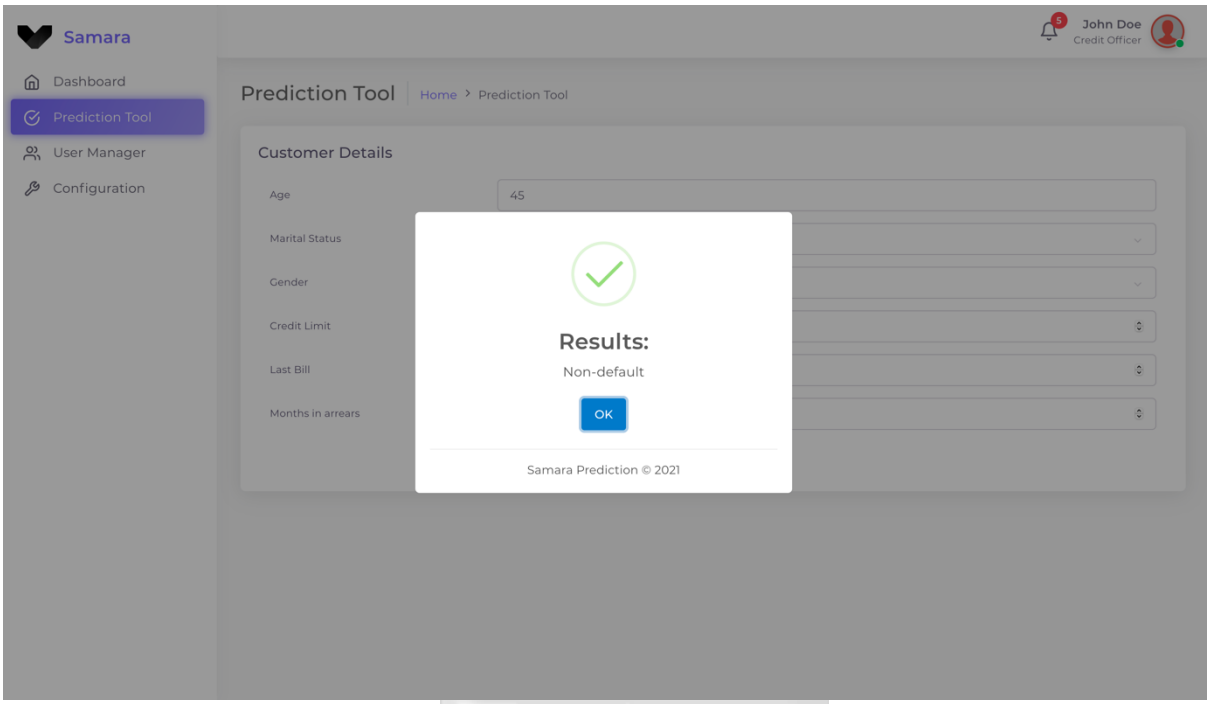

Figure 5:10 Prediction results

5.4 System Testing

The study performed the following testing techniques to ascertain the fulfillment of the functional and non-functional requirements outlined in section 4.1.2 and 4.1.3:

Table 5:2 Testing Approaches

Testing Technique	Focus
Unit testing	Check internal data structures, validation and sanitation of both input and output data.
Integrated testing	Ascertaining that the various integrated system components are working efficiently
Acceptance testing	Checking the deliverables of the system against the expected business requirements

Security testing	Eliminating security vulnerabilities and compliance to security policies and requirements for credit card data
Performance testing	Ensuring the system responds within considerable time when supplied with input data.

5.5 System Validation

When training the model, each iteration of tuning was subjected to a 10-fold cross validation to avoid overfitting and achieved a prediction accuracy 81.62% based on the evaluation criteria of a confusion matrix. The following table shows the classification results of the model:

Table 5:3 Model Classification Results

Classification	Precision Score	Recall	F1-Score
0	0.84	0.94	0.89
1	0.64	0.37	0.47

The performance of the model was further compared with other classifiers by training with the same data used in the model, and obtained the following results which showed that the XGboost model outperformed the other classifiers and therefore, is a suitable prediction model for the study:

Table 5:4 Comparison of Prediction Accuracy

Method	Prediction Accuracy
KNN	77.42

Logistic Regression	77.87
Linear Discriminant Analysis	80.97
Random Forests	81.07

Chapter 6: Discussion

6.1 Introduction

This chapter discusses the results of the study according to the objectives set out at in chapter one. The objectives of the study were to analyze the factors that influence delinquency, examine the classifiers used in predicting delinquency, develop a model for predicting delinquency and testing the performance of this model.

6.2 Determinants of pre-delinquency scoring

The study examined socio-economic, socio-demographic characteristics and behavioral characteristics of a cardholder and transaction activities relating to consumption and repayment of the credit card debt. The socio-economic characteristics of age, marital

status and age were found to be key factors that influence the propensity to repay. Behavioral characteristics such as the repayment pattern as exhibited by the pay amounts show cased that cardholders with no revolving balance and with few months in arrears were likely to repay their debt in the next billing cycle. Credit cardholders with a consecutive high number of repayment months in arrears had a high likelihood of missing payments and may point to the issue that the customer is in financial distress or is caught up in a cycle of persistent debt.

The study derived two variables from the dataset: overdrafts and credit utilization ratio. Overdrafts which were derived from the difference between the bill amounts and the credit limits were highly significant towards the probability of missing payments while the credit utilization ratio values were less significant.

The study limited to developing the model with respect to the available data and thus, some variables were not captured in the model such as direct-debit cancellations, card inactivity, cash advances, bounced payments and the type of products or services purchased by a credit cardholder. The significance of these attributes, therefore, could not be justified in this study.

6.3 Predicting pre-delinquency using XGBoost classifier algorithm

The study investigated both single and ensemble classifier techniques that fall into either statistical or machine learning algorithms that are used in predicting the likelihood of a credit cardholder missing payment. Ensemble classifiers were found to be outstanding classifiers due to their ability to improve on the weaknesses of single classifiers by trading-off variance and bias factors of the base models used. In this study, the XGBoost classifier algorithm was used to develop the model and was implemented using the python programming language. All the cardholder personal characteristics and financial information present in the dataset plus a set of derived variables including the overdraft amounts for each month and the credit utilization ratio were used to train the model. The model was optimized using the Bayesian optimization approach with a 10-cross-fold

validation performed in each evaluation. The model was then integrated into a REST API built with the Flask framework, and the functionalities exposed using graphical user interfaces built using the React framework.

6.4 Testing the performance of the model

The model achieved 81.62% prediction accuracy and outperformed a number of classifiers that were benchmarked against it. This result is suitable for prediction and an issue can adapt it and extend it to assign a probability score which can inform the business the intervention to be taken to reduce the potential risk exposure. The model can be further be extended to define a recommendation strategy which will be used for the credit management team for proactive engagement with the cardholders.

Chapter 7: Conclusion and Recommendation

7.1 Conclusions

The study had the following 4 objectives with the goal of developing a pre-delinquency model for credit cardholders. The study analyzed socio-economic, socio-demographic and transaction-oriented behavior and concluded that demographic data had lower correlation to default while the transaction-oriented behavior both in consumption and repayment contributed highly to credit cardholders defaulting. The study examined both single and ensemble classifier algorithms techniques and noted a observed a higher prediction performance in the ensemble classifiers as supported by the empirical evidence obtained by benchmarking the model with other single classifier algorithms. The main goal of the study was achieved by developing an XGBoost classifier model based on cardholders' personal characteristics and transaction-oriented behavior obtained in the financial attributes of the dataset. The model achieved 81.62% prediction accuracy and outperformed a number of single classifiers. This made this model suitable for classification purposes in this problem domain.

7.2 Recommendations

From the study, it was deduced that financial anxiety is a key contributor for credit cardholders accumulating debt and financial institutions should include key indicators of financial anxiety from transactions that span beyond the credit card account related activities such as direct debit cancellations, overall credit ratings, revolving loan debts from other accounts, etc. To minimize risk exposure, financial institutions should revise the credit limits and the minimum payment amounts depending on the repayment trends that a given credit cardholder exhibits over a given observation period. Finally, pre-delinquency scoring modeling can perform well with aggregation of a cardholder's activities across different domains, with adjustments done with respect to the current prevailing economic conditions, governing policies and frameworks, and the risk appetite of the financial institutions.

7.3 Future Work

The developed model can be extended into an expert system that evaluates early-warning pre-delinquency indicators that activates upon pattern recognition on real time. Further, the model can be aggregated with other collection scores and analytics to identify self-cures and develop appropriate contact strategies that collection teams can leverage to minimize costs and maximize collections.

References

Agarwal, S., Ambrose, B. W. (Brent W., & Liu, C. (2006). Credit Lines and Credit Utilization. Journal of Money, Credit, and Banking, 38(1), 1–22. https://doi.org/10.1353/mcb.2006.0010

Ala'raj, M., Abbod, M., & Radi, M. (2018). The applicability of credit scoring models in emerging economies: an evidence from Jordan. International Journal of Islamic and Middle Eastern Finance and Management, 11(4), 608–630. https://doi.org/10.1108/IMEFM-02-2017-0048

Barboza, G., Smith, C. and Boubacar, I. (2017). A Contribution to the Empirics of Consumers' Anxiety Behavior on and in Credit Card Repayment. Credit Card Management and Financial Literacy Among College Student. Journal of Financial Management, Markets and Institutions, (1), pp.35-66. https://doi.org/10.12831/87059

Bhattacherjee, A. (2012). Social science research principles, methods, and practices. University of South Florida. http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1002&context=oa_textbooks

Boughaci, D., & Alkhawaldeh, A. A. K. (2020). Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: A comparative study. Risk and Decision Analysis, 8(1–2), 15–24. https://doi.org/10.3233/RDA-180051

Chou, T., & Lo, M. (2018). Predicting credit card defaults with deep learning and other machine learning models. International Journal of Computer Theory and Engineering, 10(4), 105–110. https://doi.org/10.7763/ijcte.2018.v10.1208

Ciunova-Shuleska, A. (2012). The Impact of Demographic, Socio-economic and Behavioral Characteristics on Attitudes Toward Credit Cards in Macedonia. Mediterranean Journal of Social Sciences, 3(9), 199-206.

de Paula, D. A. V., Artes, R., Ayres, F., & Minardi, A. M. A. F. (2019). Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques. RAUSP Management Journal, 54(3), 321–336. https://doi.org/10.1108/RAUSP-03-2018-0003

Esgalhado, B., Higginson, M., Jacques, F., Matecsa, M. and Selandari, F. (2019). Getting a grip on bad debt: Practical steps to help utilities boost their resilience: McKinsey Digital.

Financial Conduct Authority (2017). Credit card market study: consultation on persistent debt and earlier intervention remedies.

Finlay S. (2010). Collections (Early-Stage Delinquency). The Management of Consumer Credit. Palgrave Macmillan, London.

Firafis, H. (2015). Determinants of loan repayment performance: Case study of Harari microfinance institutions. Journal of Agricultural Extension and Rural Development, 7(2), 56–64. https://doi.org/10.5897/JAERD2014.0622

Hartmann, J. (2021). Classification Using Decision Tree Ensembles, Einhorn, M., Löffler, M., de Bellis, E., Herrmann, A. and Burghartz, P. (Ed.) The Machine Age of Customer Insight, Emerald Publishing Limited, Bingley, pp. 103-117. https://doi.org.ezproxy.library.strathmore.edu/10.1108/978-1-83909-694-520211011

Kan M.P.H., Fabrigar L.R. (2017) Theory of Planned Behavior. In: Zeigler-Hill V., Shackelford T. (eds) Encyclopedia of Personality and Individual Differences. Springer, Cham. https://doi.org/10.1007/978-3-319-28099-8_1191-1

Kaur, H. and Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, Vol. ahead-of-print No. ahead-of-print. https://doi.org.ezproxy.library.strathmore.edu/10.1016/j.aci.2018.12.004

Khemais, Z., Nesrine, D., & Mohamed, M. (2016). Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression. International Journal of Economics and Finance, 8(4), 39. https://doi.org/10.5539/ijef.v8n4p39

Kiarie, F. K., Nzuki, D. M., & Gichuhi, A. W. (2013). 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.

Kumar, V. (2013). Collections scorecards and risk segmentation [Online]. Available at https://www.experian.co.uk/blogs/latest-thinking/decisions-and-credit-risk/collections-scorecards-and-risk-segmentation/ (Accessed: 31 May 2020)

Lada M & Wejer-Kudelko M (2018). Success factors and barriers to the effective debt collection process. Scientific Works of the Wrocław University of Economics, (51`5), pp. 147-155.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lin, L., Revindo, M.D., Gan, C. and Cohen, D.A. (2019), "Determinants of credit card spending and debt of Chinese consumers", International Journal of Bank Marketing, Vol. 37 No. 2, pp. 545-564.

Ming-Yen Teoh, W., Chong, S., & Mid Yong, S. (2013). Exploring the factors influencing credit card spending behavior among Malaysians. International Journal of Bank Marketing, 31(6), 481–500. https://doi.org/10.1108/IJBM-04-2013-0037

Mittal, S., Gupta, P., & Jain, K. (2011). Neural network credit scoring model for micro enterprise financing in India. Qualitative Research in Financial Markets, 3(3), 224–242. https://doi.org/10.1108/17554171111176921

Onay, C., & Öztürk, E. (2018). A review of credit scoring research in the age of Big Data. Journal of Financial Regulation and Compliance, 26(3), 382–405. https://doi.org/10.1108/JFRC-06-2017-0054

Régis, D. E., & Artes, R. (2015). Using multi-state markov models to identify credit card risk. Production, 26(2), 330–344. https://doi.org/10.1590/0103-6513.160814

Shapiro, G. K., & Burchell, B. J. (2012). Measuring financial anxiety. Journal of Neuroscience, Psychology, and Economics, 5(2), 92–103. https://doi.org/10.1037/a0027647

Singh, S., Rylander, D. H., & Mims, T. C. (2018). Understanding credit card payment behavior among college students. Journal of Financial Services Marketing, 23(1), 38–49. https://doi.org/10.1057/s41264-018-0042-0

Stavins, J. (2020). Credit card debt and consumer payment choice: What can we learn from credit bureau data? Journal of Financial Services Research, 58(1), 59-90. https://doi.org/10.1007/s10693-019-00330-8

Taneta- Skwiercz, D. (2018). Sustainable development indicators - Poland against the background of European Union countries. Scientific Works of the Wrocław University of Economics, (516), pp.121–132.

Team, S. (2016). Debt Management and Collection Analytics (Post Delinquency Vintage) | Scoredata. [online] Scoredata.com. Available at: https://scoredata.com/debt-

management-and-collection-analytics-post-delinquency-vintage/ [Accessed 5 June 2020].

Teng, H.-W., & Lee, M. (2019). Estimation procedures of using five Alternative machine learning methods for predicting credit Card Default. Review of Pacific Basin Financial Markets and Policies, 22(03), 1950021. https://doi.org/10.1142/s0219091519500218

Thomas, L., Edelman, D. and Crook, J. (2017). Credit Scoring and Its Applications. 2nd ed. Philadelphia: Siam, ISBN 978-1-611-97455-3, pp. 157-177.

Tsai, C.-F. and Hung, C. (2014). Modeling credit scoring using neural network ensembles, Kybernetes, Vol. 43 No. 7, pp. 1114-1123. https://doi.org.ezproxy.library.strathmore.edu/10.1108/K-01-2014-0016

Varma, R. (2015). Bolstering Credit with Pre-Delinquency Management. [online] BAI. Available at: https://www.bai.org/banking-strategies/article-detail/bolstering-credit-with-pre-delinquency-management/ [Accessed 5 June 2020].

Wong, K.Y. and Lynn, M. (2019), "Credit card cue effect: How mere exposure to credit card cues promotes consumers' perceived financial well-being and spending", International Journal of Bank Marketing, Vol. 38 No. 2, pp. 368-383.

Zainudin, R., Mahdzan, N. S., & Yeap, M.-Y. (2019). Determinants of credit card misuse among Gen Y consumers in urban Malaysia. International Journal of Bank Marketing, 37(5), 1350–1370. https://doi.org/10.1108/IJBM-08-2018-0215

Zhanga, C., & Ma, Y. (2012). Ensemble machine learning: Methods and applications. Springer.

Appendix

Appendix A: Strathmore University Institutional Ethics Review Committee Approval

Strathmore UNIVERSITY

28th April 2021

Mr Kisengese, Antony
antony.kisengese@strathmore.edu

Dear Mr Kisengese,

RE: A Model for Predicting Pre-Delinquency Based on Transaction-Oriented Cardholder Behaviour Using Random Forests

This is to inform you that SU-IERC has reviewed and **approved** your above **master's** research proposal. Your application reference number is **SU-IERC0936/20.** The approval period is **28th April 2021 to 27th April 2022.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and also obtain other clearances needed

Yours sincerely,

for: Dr Virginia Gichuru,
Secretary; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE (SU-IERC)
28 Apr 2021
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Cc: Prof Fred Were,
Chairperson; SU-IERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu