University of Cape Town

ECO 5016W

Minor Dissertation in Financial Technology

Using Supervised Learning Methods to Credit Score Informal Merchants

Student:

Derick Kazimoto

Student Number:

KZMDER001

Supervisor:

Dr. Şebnem Er

Co-Supervisor:

A/Prof. Co-Pierre Georg

August 21, 2023

Contents

List of Tables

List of Figures

Declaration

I, Derick Kazimoto, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: . . . .

Date: 21-08-2023 . . . .

Acknowledgements

Throughout my masters degree I have received a great deal of support and assistance. I would like to first thank the support of the South African Reserve Bank (SARB) for funding my masters and the SARB Chair in Financial Stability Studies, A/Professor Co-Pierre Georg, for believing in my abilities and for the invaluable critical thinking sessions incorporated in your lessons.

I would like to thank my supervisor, Dr. Şebnem Er, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

I would also like to thank Nomanini for their wonderful support in this study and giving me the opportunity to undertake this research. I would particularly like to single out Dale Humby, the Chief Technical Officer of Nomanini. Dale, I want to thank you for your patient support and for the opportunity I was given to undertake my research with Nomanini.

I would like to acknowledge the BFA Global and Cambridge Centre of Alternative Finance, Sara Coupe, for placing me in the Catalyst Fund Fintech Program with Spoon Money. My six (6) month experience working with Nicky Swartz and Lorna McLaren contributed to my better understanding of how to extract value from machine learning models to business decision making tools. This learning experience was important and incorporated to my research.

I would like to thank Zanele, Sophia, Michele, Zipho and Michael for your support that helped develop my resilience in tough times and for the stimulating discussions

as well as happy distractions to rest my mind outside of my masters.

In addition, I would like to thank my parents and family for their endless support, wise counsel and sympathetic ear. You are always there for me.

Abstract

Access to working capital is a significant challenge for the informal retail sector, where better financial products are not easily accessible. This study aims to address this issue by developing data-driven credit scoring models for informal merchants using supervised learning methods. The study uses data collected from Nomanini, a financial technology company that facilitates loans to informal merchants in Lesotho. The objective of the study is to help Nomanini develop accurate credit scoring models to reduce the risk of lending money to potential defaulters.

Logistic regression and support vector machines were used as supervised learning methods to predict the default behavior of merchants. Six (6) logistic regression models and twelve (12) support vector machine models were evaluated based on their default predictive power. The best-performing model was a logistic regression model that used a merchant's credit history as the only feature, resulting in a Gini coefficient of 0.6143.

The study's findings can help Nomanini determine the creditworthiness of merchants more accurately and reduce the risk of lending money to defaulters. This, in turn, can help increase access to working capital to support and grow small businesses in the informal sector. In conclusion, the study highlights the potential of using supervised learning methods to develop credit scoring models for informal merchants and contribute towards reducing the financial gap in the informal African economy.

Keywords: credit scorecards, credit scoring, default, informal merchants, logistic regression, support vector machines, Gini coefficient.

1 Introduction

1.1 Research Motivation

This study aims to develop credit scoring models for informal merchants in the informal market by experimenting with different supervised learning methods to predict their default risk. The data used in this study was collected from Nomanini, a financial technology company that facilitates loans to informal merchants in Lesotho. The models developed in this study can help financial institutions make better-informed decisions on providing credit to informal merchants.

1.2 Background of Study

Borrowing and lending have a long history associated with human behavior [27]. Despite the fact that credit has been around since 2000 BC or earlier, credit scoring has only been around for about six decades. The term credit scoring refers to the process of assessing an applicant's risk of defaulting on a financial obligation [14]. Financial institutions create credit scorecards based on information gathered from credit applicants [15] [27].

Credit scoring has recently become much easier to use due to advances in computing technology, which has increased its applications in a variety of fields. These advanced technologies employ improved classification techniques which ensure that only those applicants with a small probability of defaulting are offered loans [1].

1.3 Objectives of Research

Data is an essential component in assisting financial institutions to make credit decisions on credit applicants. In this study, anonymized transactional and loan data were used to achieve the following research objectives:

- Implement logistic regression and support vector machines and evaluate their performance at predicting default of credit applicants.
- Compare the performance of the logistic regression models and support vector machines.
- Recommend the best model to use to predict the default behaviour of credit applicants.

1.4 Assumptions of the Study

Despite the fact that credit scorecards have a limited lifetime due to changing economic conditions, the study assumes that applicants' default behavior will remain consistent in the future [3]. Furthermore, the study assumes that the number of defaulters in the dataset is sufficient to predict customer default risk. The problem with this assumption is that customers sometimes default for reasons unknown to the financial institution, posing a risk to the performance of credit scoring models.

2 Literature review

2.1 Credit scoring in informal markets

Financial institutions offer critical support for Micro, Small and Medium Enterprises (MSMEs) to grow, and to carry out everyday transactions more easily to finance expansion. The financial services required by MSMEs differ significantly according to their size, degree of formality, and growth stage – but mostly require is some form of transactional banking or payment product to help them transact with their suppliers, customers and staff; and credit to help sustain or grow the business [7].

Credit scorecards are becoming more popular among banks as a way to assess risk and determine creditworthiness for small businesses with higher risk profiles. This allows for a quicker risk assessment process and can reduce the need for traditional collateral, as working capital financing is provided without the need for security [7]. However, the use of scorecards is unlikely to fill the significant financing gap if a large segment of the MSME sector remains informal and cash-based, as banks and other financial institutions require data over a time period to apply this method [7].

2.1.1 Nomanini

Nomanini is a pioneering Financial Technology platform that is headquartered in South Africa and founded in 2011 to innovate for the informal retail eco-system. It connects the merchants and distributors to each other and to global service providers who integrate payments, working capital and data analytics to unlock the latent potential of African's economy. Nomanini operates by connecting any mobile device into a retail point of sale solution for informal merchants.

The wallet allows merchants to provide a broad range of services such as digital banking, mobile, utility and entertainment services to their customers and therefore, boosts the merchants' competitiveness. The platforms enable digital service providers to increase the scale of their business. Based on the real-time insights generated through transactional data, distributors can improve sales by gaining a single view of their merchant network and ensure inventory is where it is needed most. Also, distributors can begin to accept payments for goods electronically and thus, eliminating the risk and inefficiency of collecting cash.

Moreover, by utilizing data analytics, Nomanini is able to offer working capital loans to informal merchants through their distributors, enabling them to expand their businesses. This leads to an increase in the volume of goods and services traded, while also reducing operational friction, and ultimately resulting in higher profits for all parties involved.

In a pilot project, to improve the operational environment of informal merchants to conduct business, by the Financial Inclusion of Business Runaways (FIBR) and it was found that access to working capital is the most significant barrier to growth in the informal retail sector.

Therefore, it is important to adopt a merchant-focused approach in addressing problems associated with reliance on cash, inadequate working capital, and high transactional costs. The Nomanini Fintech platform, a managed cloud solution, facilitates connectivity between service providers, such as banks and Fast Moving Consumer Goods (FMCG) distributors, and informal economy merchants, promoting unity within the landscape. The platform aims to resolve these challenges in the following ways:

- Nomanini helps reduce the usage of cash by turning merchants' mobile devices into a retail point-of-sale (POS) device connected to an inter-operable merchant wallet. This enables them to facilitate a range of electronic payments from multiple service providers without investing in further infrastructure.
- The data generated from their mobile POS will help nurture trust between merchants and distributors. Additionally, it will unlock credit flows from banks as a result of reduced risks that will provide traders access to working capital.

To unlock the latent potential of Africa's informal economy, merchant's need to access financial offerings that will meet their real needs and challenges [20]. With Nomanini's platform, merchants are able to interact with banks and distributors as a business looking for a loan and not an individual looking for a microloan. The shift in the notion helps merchants to elevate their success and growth. This optimises relationships between merchants, distributors and manufacturers by unlocking value and scale throughout the value chain in a digital manner. Hence, by enabling merchants to access working capital, invest in their businesses and reduce the reliance on cash, Nomanini's solution drives growth and reduces friction for everyone in the value chain [19]. Therefore, through a strength in partnerships between banks and financial technology companies, a provision of market-loan product fit can be achieved to provide working capital to the informal sector.

2.1.2 Financial technology and the informal retail economy in Africa

According to the International Monetary Fund (IMF), Sub-Saharan Africa has the second largest informal economy in the world after Latin America and the Caribbean. Even though the Sub-Saharan informal economy accounts for 38% of the region's

GDP, the informal traders that drive this market are under served and excluded from the formal financial system [18]. Hence, this problem provides an opportunity for innovative financial technological solutions that will address the pain points of the informal sector, improve efficiencies and boost economic growth.

Despite the growth in mobile money and mobile payments in Sub Saharan Africa, cash is still king. The migration from cash in Africa is likely to take decades and this reliance of cash as a medium of exchange prevents the formal financial services to have digital data that they can use to develop credit scoring models. Hence, the informal sector lack the working capital required to invest in goods and services that will grow their businesses [19].

Apart from the challenge of reliance of cash and lack of working capital, the informal sector faces a challenge of a fragmented value chain which lacks a uniform payments infrastructure that banks, distributors and other members of the eco-system can use to have a view of the informal market and the merchant's needs. The fragmentation of markets, dividing them between digital, financially included, urban areas and analogue, financially excluded, rural areas, results in complex efforts to achieve scalability and decrease friction in the supply chain. This division may also impact the availability of data for credit scoring purposes.

Innovative financial technology solutions can however provide a bridge between the informal economy and formal financial services, by rewriting the rules of informal trade in Africa. Through connecting informal merchants to digitised financial services, offers the opportunity to improve efficiencies across the value chain, and ultimately boost economic growth.

2.2 Building credit scoring models

Credit score models help financial institutions to serve low-income customers and grow their portfolios [28]. The decision-making process can be judgmental, statistical, or a combination of both [24]. Judgemental scoring depends on an expert to provide a qualitative judgment, whilst a statistical approach depends on quantified characteristics, a set of rules, and statistical techniques to forecast risk as a probability. Statistical scoring models are important where lenders need to perform a large volume of credit assessments for loan amounts that are relatively low and for retail credit for individuals and small businesses [28]. As shown in Table 1, the two approaches complement each other with different benefits and challenges [28].

Dimension	**Judgemental Scoring**	**Statistical Scoring**
Source of knowledge	Experience of credit expert	Quantified portfolio history
Consistency of process	Varies	Identical loans scored identically
Explicitness of process	Evaluation guidelines in office	Mathematical rules to quantify risk
Process & Product	Qualitative classification	Quantitative classification
Ease of acceptance	Common	Uncommon
Process of implementation	Lengthy training for credit experts	Lengthy training for stakeholders
Vulnerability to abuse	High	Low
Flexibility	Wide application	Limited to specific risk
Knowledge of trade-offs	Based on experience	Derived from tests

Table 1: Comparison between judgemental scoring and statistical scoring [24].

2.2.1 Data sources and quality

In financial institutions data is collected from multiple sources. The ability to use data for analysis and obtaining of actionable insights depends on the quality of the data. Hence, it is important to understand where the data comes from and how it is captured.

Different data fields have varied consistency in collection which affects the expected reliability in predictive analytics. Therefore, the following are the description of the different types of data and their reliability:

- Transactional (High reliability) - When stored consistently, transactional data have a high reliability. These data points include account deposits, withdrawals, loan payments, bill payments and many more. They provide a history and objective of a customer's actual behaviour and economic activity [28].
- Documentary (High reliability) - Identity and demographic document data which are verified by the government. For example national ID cards, driving licenses, passports and any other relevant government related identity card [28].
- Collected from devices (High reliability) - The data from the devices can be as reliable as the transactional data [28].
- Psychometric (Above average) - These are self-reported tests given to customers in form of tests or questionnaires. The data reliability depends on the quality of the tests [28].
- Collected from staff (Average) - These data may be affected by human error which include judgement, experience of the person collecting it and the work style [28].
- Self reported (Below average) - Data reported by customers tend to have low reliability since they tailor the responses to maximize their chances of being approved [28].

2.2.2 Credit scoring models

From the different sources of data, a credit scoring model is built to predict default. The key assumption used is that the past behaviour of customer repayments will resemble the future in terms of default risk. Hence by definition, default risk is the probability that the borrower will fail to pay the loan [3].

The statistical approach is based on statistical analysis of historical data that finds the optimal multivariate relationship between a customer's characteristics and default behaviour [4]. A scorecard of multivariate correlation of inputs such as age, marital status, income, savings amount and a target variable reflecting the risk of default. Each input will be assigned a score that will be added and then compared to a threshold that determines the credit quality of the applicant. Since statistical credit score cards are mathematical formulas, they can be easily programmed and evaluated in a fast way. This makes it easier to make credit decisions in an online setting where decisions need to be made as fast as possible. Also, another benefit of using statistical methods is consistency. This approach uses a formal data-based modelling that eliminates the biases which might arise from subjective decisions made by credit experts. The method will always evaluate the same inputs in the same way [3].

The two approaches used in building statistical credit scoring models are: application and behavioural scoring described below.

- **Application Scoring** is the statistical credit approach of determining a credit score that reflects the default risk of a customer upon loan application [3].
- **Behavioural Scoring** is another statistical approach that analyses the be-

haviour of existing credit customers. After credit has been granted, lenders use behavioural scoring to assess the likelihood of default occurring during some specific outcome period [3].

Most credit scoring models are built using proven classification methods such as logistic regression and decision trees to estimate the probability of default [28]. Logistic regression is the most popular technique in the industry as it provides a continuous range of scores between 0 and 1. However, decision trees are mostly used during data pre-processing for feature selection, categorization or segmentation [3].

Logistic regression was first used in financial risk by Ohlson, where he applied it to predict bankruptcy [21]. Thereafter, Wiginton [29] became the first person to specifically use logistic regression to predict credit default with the objective of comparing it with linear discriminant analysis models, that were common at the time. Currently, logistic regression is the most popular classification technique used in credit default prediction because of its simplicity and interpretability [3]. Logistic regression is a generalized linear model that predicts discrete outcomes. The response variable in the binary logistic regression is either 1 or 0, with a probability of θ and $1-\theta$ respectively. For credit risk analytics, a random variable Y_i takes value of 1 if the loan is defaulted and the value of 0 if the loan is not defaulted $(i = 1, \ldots, n)$. Hence, the probability of defaulting is defined by $\theta = \Pr(Y_i = 1|X)$ and the probability not defaulting by $1 - \theta = 1 - \Pr(Y_i = 1|X)$ [30].

The relationship between the response and the independent variables (X) are described by the linear logit transformation as follows:

$$logit\,(\theta) = log_e\left[\frac{\theta}{1-\theta}\right] = \alpha + \beta^T X \tag{1}$$

where α is the intercept and $\beta = (\beta_1, \ldots, \beta_p)^T$ is the vector of slope parameters for the independent variables ($p = 1, \ldots, k$). By this transformation $\theta = \Pr(Y_i = 1|X)$ is defined as

$$\theta = \frac{e^{\alpha+\beta^T X}}{1 + e^{\alpha+\beta^T X}} \tag{2}$$

where $0 \leq \theta \leq 1$.

The unknown β parameters are estimated using likelihood function as follows 3[16]:

$$L(\beta) = \prod_{i=1}^{n} [p_i(X)]^{y_i}[1 - p_i(X)]^{(1-y_i)} \tag{3}$$

The optimum solution is obtained by maximizing the likelihood function with respect to the β parameters where the likelihood function is transformed using a log transformation:

$$lnL = \sum_{i=1}^{n} Y_i ln\left(\frac{e^{\alpha+\beta^T X_i}}{1 + e^{\alpha+\beta^T X_i}}\right) + \sum_{i=1}^{n}(1 - Y_i)\, ln\left(\frac{1}{1 + e^{\alpha+\beta^T X_i}}\right) \tag{4}$$

This modelling technique provides a linear combination of independent variables X_p with coefficients β_p which estimates the likelihood of a loan being defaulted or not [5]. The formula in Equation 2 estimates the probability of default [30]. If the probability is greater than or equal to 0.5, the loan is grouped into default, and if not it is grouped into not defaulted.

Support Vector Machines (SVMs) are a popular supervised learning method that seek to find an optimal hyperplane that constructs a maximum margin separator that creates a decision boundary with the largest distance to example points as shown in Equation 5[11]. Where w is the normal vector to the hyperplane and b is the bias term. SVMs aim to find the hyperplane that maximizes the margin, subject to the constraint that all training examples are classified correctly. This can be formulated as a quadratic programming problem, which can be solved using standard optimization techniques [8].

$$w \cdot x + b = 0 \tag{5}$$

In the soft-margin SVM formulation, slack variables ξ_i are introduced to allow for the misclassification of training examples, and a hyperparameter C is introduced to control the trade-off between maximizing the margin and minimizing the classification error on the training data [8]. The soft-margin SVM formulation can handle linearly non-separable data, making it a more flexible variant of the SVM. The optimization problem can be formulated as follows:

$$\min_{w,b,\xi} \frac{1}{2}|w|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{subject to: } y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0, i = 1, 2, ..., n \tag{6}$$

Where n is the number of examples in the training dataset, (x_i, y_i) is the i-th example and its corresponding binary class label, and C is a hyperparameter that determines the penalty for misclassifying training examples. A small value of C will produce a wider margin and allow more misclassifications, while a large value of C will produce a narrower margin and fewer misclassifications.

For example, in Figure 1 (a) the three candidate decision boundaries correctly classify the classification problem of black and white circles. This is the case for logistic regression as it is optimized to minimize the empirical loss of the training data and thus, all three separators are equally as good. However, SVMs are optimized to minimize expected generalization loss with a maximum margin separator as shown in Figure 1 (b).

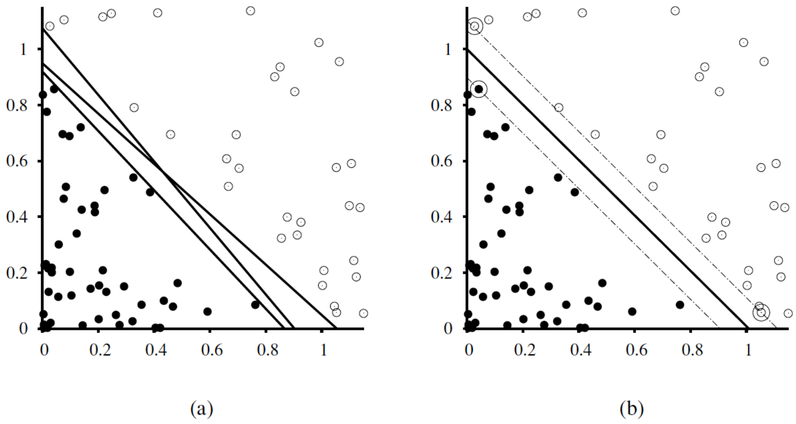

Figure 1: Figure 1(a) shows two classes of points (black and white circles) and three candidate linear separators. Figure 1(b) shows the maximum margin separator (heavy line) is at the midpoint of the margin between dashed lines. The support vectors are the points with large circles on the dashed lines as the closest examples to the separator[23].

The kernel trick is used in SVMs to transform the input space into a higher dimensional feature space, allowing nonlinear decision boundaries to be drawn in the original space. Two commonly used kernel functions are the linear and radial basis function (RBF) kernels. The linear kernel is suitable for linearly separable data while the RBF kernel is used for nonlinearly separable data [26]. The RBF kernel has a hyperparameter γ (Gamma) that determines the shape of the decision boundary. A small value of γ (Gamma) will produce a smooth decision boundary and may lead to underfitting, while a large value of γ (Gamma) will produce a complex decision boundary and may lead to overfitting. To find the optimal values for C and γ (Gamma), a grid search can be used to test different combinations of values and evaluate their performance using a validation set. Alternatively, Bayesian optimization can be used to iteratively search for the optimal values based on the model's performance [25].

For example Figure2(a), the data points are not linearly separable in a two-dimensional space but when mapped into a higher dimension space as shown in Figure2(b), they become linearly separable.

Overall, SVMs are a powerful tool for building credit scoring models as they can handle both linearly and nonlinearly separable data and can be optimized to balance between maximizing the margin and minimizing classification error.

Other classification methods include decision trees, which are recursive partitioning algorithms that develop a tree-like structure representing patterns in an underlying dataset [9]. Since the 1980s, decision trees have been employed for developing credit score models; however, their limitations stem from instability caused by fluctuations in the data sample, potentially leading to significant variations in feature choices at each node and the classifications assigned to instances [6].

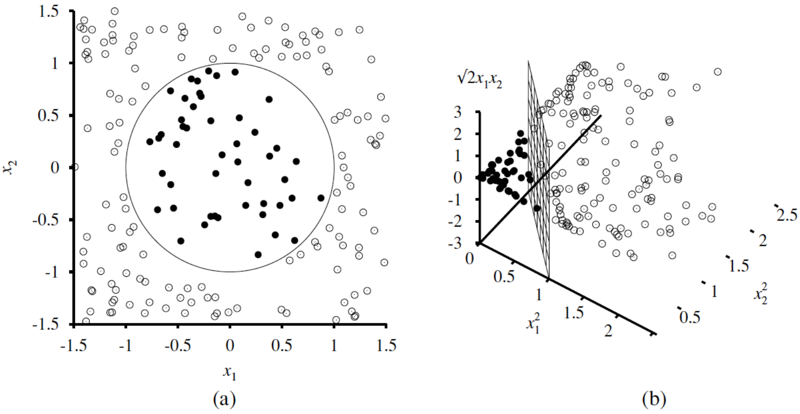

Figure 2: Figure 2(a) shows a two-dimensional training dataset of black and white circles with a decision boundary of $x_1^2 + x_2^2 \leqslant 1$. Figure 2(b) Shows the same training dataset after mapping it into a three-dimensional space $(x_1^2, x_2^2, \sqrt{2}x_1x_2)$. The circular boundary in Figure 2(a) is transformed into a linear decision boundary in Figure 2(b) [23].

In addition to decision trees, other classification models which have been developed to build credit scorecards, including discriminant analysis, neural networks, ensemble methods such as bagging, boosting, and random forests [2]. While these other techniques are considered to be more powerful and may provide better prediction results, they often produce very complex models that make them less useful in building credit scoring models where interpretability is a key concern [17]. According to Hand (2006), the potential performance improvements attained by complex models are small compared to the predictive power of simple models and are often offset by other sources of uncertainty that are exacerbated by the added complexity [13].

2.2.3 Evaluating a credit score model

Once a model is developed, different alternatives are used to choose the one that yields the highest predictive power. The model will be evaluated on a different data sample (test set) which is about 30% of the total data sample. The following are the methods used to evaluate the performance of a credit score card [28]:

- Confusion Matrix
- Receiving Operating Characteristic Curve, Area under the Curve and Gini Coefficient
- Kolmogorov-Smirnov Test

Confusion matrix is a table layout that is used to evaluate the performance of a classification algorithm. As shown in Table 2, 0 represents a not default status and 1 represents a default status, the matrix shows two types of errors that may be committed consequences for the business:

- False Positive - Predicting that a loan applicant will default (1) when in fact the applicant is actually a good loan applicant (0). The potential loss to the business is interest and loan fees.
- False Negative - Predicting that a loan applicant is good (0) when in fact the applicant is actually a bad loan applicant (1). The potential loss for the business is profits.

As for the correct predictions the matrix in Table 2 shows two types of good predictions that benefit the business as follows:

- True Negative - Predicting a good loan applicant (0) when in fact the applicant is actually a good loan applicant (0). The potential benefits for the business is interests paid and the recovery of the loan costs and amount.
- True Positive - Predicting a bad loan applicant (1) when in fact the applicant

is a bad loan applicant (1). The potential benefit gained for the business is the avoidance of losses from loan costs and amounts.

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

Table 2: Confusion Matrix

There are various performance metrics which can be obtained from the confusion matrix such as:

- Accuracy
- Precision
- Specificity
- Recall/ Sensitivity
- F1 Score

Accuracy metric is ratio of the the total number of correctly predicted samples to the total samples. As shown in Equation 7, the total number of true positive and true negatives are divided by the total samples. This performance metric measures how many samples were correctly predicted by the model.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

Precision metric is the ratio of the true positives to the total number of positive predicted samples (both true positives and false positives) in the dataset. This metric measures how many predicted positives are actually true. Equation 8 shows the formula that calculates the precision performance metric of a classification algorithm.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Specificity metric is the ratio of the true negatives to the total actual negative samples (true negatives and false positives) by the classification models. This metric measures how many actual negatives were predicted correctly. Equation 9 shows the formula for calculates the recall performance metric of a classification algorithm.

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

Recall/ Sensitivity metric is the ratio of the true positives to the total actual positive samples (true positives and false negatives) by the classification models. This metric measures how many actual positives were predicted correctly. Equation 10 shows the formula for calculates the recall performance metric of a classification algorithm.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

F-Score metric combines the precision and recall metric performances as shown in Equation 11. This performance metric is used to obtain a balance between precision and recall.

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

Receiving operating characteristic (ROC) curve, area under the curve (AUC) and Gini coefficient are good metrics to use for classification model evaluations. The ROC Curve is used to simultaneously show the two types of errors for all possible cut-off probability thresholds of classification [16]. The curve plots the true positive rates versus the negative positive rates for all cut-offs and thresholds. The true positive rate which is also known as recall or sensitivity is the ability to correctly identify actual positives (defaulters). As shown in Equation 13, the negative positive rates measure the proportion of false positives that were incorrectly classified as defaulters to the sum of false positives and true negatives.

$$TruePositiveRate = \frac{TP}{TP + FN} \tag{12}$$

$$FalsePositiveRate = \frac{FP}{FP + TN} \tag{13}$$

In Figure 3, the dotted line represents the performance of the model whose predictive power is equivalent to random guessing [28]. The AUC measures the overall performance of the classifier of all the thresholds as an area under the ROC curve [16]. The area is measured as a percentage of the box in green in Figure 3. The Gini coefficient is the linear transformation of the AUC, that is a scale of the predictive power of the model. Please see Equation 14, for the Gini coefficient.

The AUC is obtained by integrating the ROC curve which is lower bounded by 0.5 (the dotted line). Hence, the AUC is at a maximum of 1 and a minimum of 0.5 and the Gini coefficient is at a maximum of 1 and a minimum of 0. Therefore, a Gini

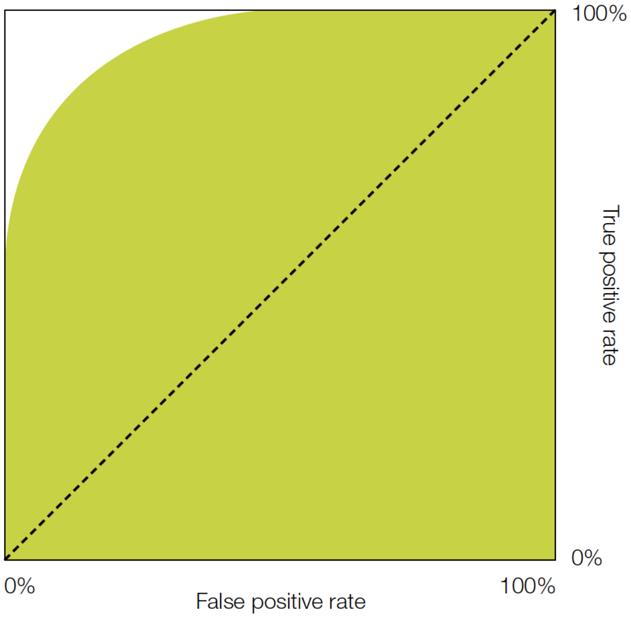

Figure 3: Example of a ROC curve [28].

of 0 and AUC of 0.5 is a random prediction, and a Gini of 1 and an AUC of 1 is a perfect prediction [28].

$$Gini = 2 \times AUC - 1 \tag{14}$$

Kolmogorov-Smirnov (KS) Test measures the maximum vertical separation between two cumulative curves (defaulters and non-defaulters) in a credit score card [28]. The difference between the curves is known as the KS score. A high KS Score indicates a high quality of the credit score card in discriminating defaulters and non-defaulters [12].

For example in Figure 4, the maximum difference in the accumulated rates happen at 67% of defaulters and 23% of non-defaulters, resulting in a KS score value of 44%.

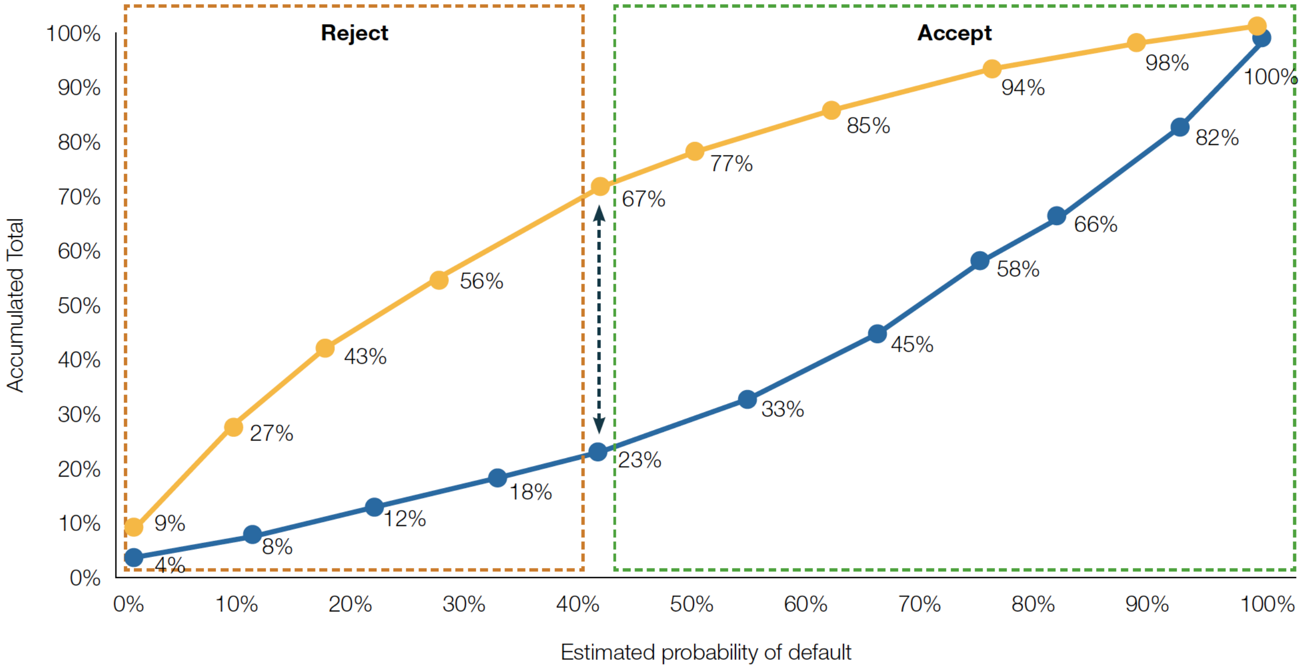

Figure 4: Example of a Kolmogorov-Smimov curve [28].

Therefore, accepting applicants at that point, the organization will be accepting 77% of non-defaulters and 33% of defaulters.

2.3 Summary of literature

Most financial institutions fail to give credit to the informal sector due to cash based operations. In practice, lenders require data over a time period to use as a benchmark to make credit decisions. This chapter discussed the financing gap that exists and how innovative technological solutions can be used to solve the problem.

With the availability of data through technological platforms such as Nomanini, the chapter reviews the literature of how credit scoring models can be built and enable financial institutions to provide credit in the informal sector. Thus, the next chapter will summarize how data was extracted from the different sources to create the datasets used for the credit scoring models.

3 Methodology

3.1 Data extraction

The dataset comprises of anonymized transaction and loan data from Nomanini's operations in Lesotho spanning from April to November 2020. During this period, merchants requested for loans from the platform with a defined repayment period of up to 9 AM SAST (collection time) on the day after the loan request. It was required for merchants to settle the loan within this time frame. Failing to do so led to the loan being classified as either "Delinquent" or "Default".

A merchant is categorized as a "Default" borrower if they completely failed to make any loan payment or if they paid the loan after more than 24 hours from the collection time. On the other hand, a merchant is considered a "Delinquent" if they managed to repay the loan within 24 hours after the collection time.

It is important to note that the final sample size after merging the data was 62 observations, which is small. Therefore, it is likely that the predictions and results may be affected due to the small sample size.

3.1.1 Transaction data

The transaction dataset consists of 793,372 observations with features of the following description:

- Id - A unique alphanumeric code identifying the transaction.
- Account id - A unique alphanumeric code identifying a merchant on the platform.

- Time - The time in which the transcaction was done.
- Kind - The type of transaction done by the merchant on the platform. The following are four kinds of transactions:–
    - Airtime Pin - This transaction type occurs when a merchant buys airtime through the platform.
    - Loan Disbursement - This transaction type occurs when a merchant is loaned out money.
    - Merchant Deposit - This transaction type occurs when a merchant makes a deposit on the platform.
    - Loan Repayment - This transaction type occurs when a merchant repays a loan through the platform.
- Amount - The monetary value of the transaction.
- Money transaction id - A unique alphanumeric code identifying a transaction.

The transaction data of a merchant was preprocessed by aggregating the total number of the kind of transaction. As a result, the transactions dataset was transformed to include the following features:

- Account id - An alphanumeric code representing the id of a merchant. It is also known as merchant id in other datasets.
- Number of airtime pins - The number of times a merchant bought airtime on the platform.
- Number of loan disbursements - The number of times merchant was loaned

out money.

- Number of merchant deposits - The number of times a merchant made a deposit to the platform.
- Number of loan repayments - The number of times a merchant repaid a loan through the platform.
- Amount - The total amount in monetary value of the transactions performed by a merchant.

3.1.2 Loan data

Loan information were extracted from following three datasets;

- Loans dataset - A dataset that consist of the loans taken by merchants.
- Offers dataset - A dataset that comprise the details of the loan products available on the platform.
- Merchant-offer dataset - A dataset that consist of the loan product details taken by a merchant.

3.1.2.1 Loans dataset

The loans dataset consist of a total of 966 observations of loans taken from May 12, 2020 to November 26, 2020 with the following feature description:

- Id - A unique alphanumeric code identifying the loan.
- Merchant offer id - A unique alphanumeric code used to extract details of a merchant's loans offer from the offer and merchant-offer datasets.

- Opened time - The time that the loan was requested by the merchant to the platform.
- Closed time - The time that the loan request was accepted back to the merchant.
- Period start - The beginning period of the loan.
- Period end - The end period of the loan.
- Collection time - The time of loan collection from the merchant's wallet. The collection time is 9 AM SAST, the day after loan period time ends.
- Status - The loan's status is either open or closed. Closed status indicates that the loan has completed its period, whereas open status indicates that the loan is still active.
- Repayment status - This feature describes whether a loan is paid back in full or not. The repayment status is categorized in the following ways:
  - Ok - The merchant paid the loan within the period length of days.
  - Delinquent - The merchant paid the loan within an extra day of the period length of days.
  - Default - The merchant either paid the loan after 24 hours of collection time or did not pay at all.

The features of the loans data set were transformed and for each merchant information on the number of defaulters, delinquents, and paid-back loans was aggregated in order to extract valuable insights that will determine the credit behavior of merchants. The transformed data had the following features:

- Account id - This is an alphanumeric code representing the id of a merchant. It is also known as merchant id in other datasets.
- Number of paid loans - The number of times a merchant repaid back a loan on time.
- Number of delinquent loans - The number of times a merchant repaid back a loan within 24 hours after collection time.
- Number of default loans - The number of times a merchant was not able to repay a loan within 24 hours after collection time.

3.1.2.2 Offers dataset

The offers dataset contains all of the loan products available on the Nomanini platform, each with the following feature description:

- Id - A unique alphanumeric code identifying a loan product.
- Title - The name of a loan product.
- Valid from - The time from which the loan product was available on the platform.
- Valid until - The time that the loan product was discontinued to be offered on the platform.
- Amount - The monetary value of the loan product.
- Fee - The cost the merchant incurs taking a loan.
- Period length of days - The period time that the loan product has to be paid.

- Category id - A unique id of the provider of the loan product.

3.1.2.3 Merchant-offer dataset

This dataset consist of the loan offer products taken up by merchants on the platform that have the following feature description:

- Id - This is the unique alphanumeric code that is also available in loan's dataset as the merchant offer id.
- Merchant id - The id of the merchant taking the loan product. Also known as account id.
- Offer id - This is the unique alphanumeric code that is also available in the offer's dataset that represents a specific loan product.
- Valid from - The time period from when the merchant took the loan.
- Valid until - The time period to when the merchant paid/defaulted the loan.

3.2 Merging the datasets

To build predictive statistical models using supervised learning, the various datasets are combined into a single dataset that includes both loan and transactional features of a merchant. As a result, the transactional, loans, offers, and merchant-offer datasets have been merged into a single dataset, yielding the following features:

- Account id
- Number of airtime pins
- Number of loan disbursements

- Number of merchant deposit
- Number of loan repayments
- Amount
- Number of paid loans
- Number of delinquent loans
- Number of default loans
- Default status

The merged dataset has 62 observations with predictive features based on transactional and loan data from April 2020 to October 2020, with the target outcome being the merchant's default status in November 2020. It is important to note that only closed loans were used in the study to ensure that customers who would not have had a chance to repay or default were excluded from the analysis.

3.3 Feature engineering

In the field of predictive modeling, feature engineering is an important process that involves the extraction of relevant and informative features from raw data to enhance the model's performance. The ultimate objective of feature engineering is to identify the most significant features that are likely to impact the model's accuracy. In the context of predicting merchant defaults, two significant features have been identified, namely the "Average Amount Per Transaction (AAPT)" and "Credit history".

3.3.1 Average Amount Per Transaction (AAPT) feature

As shown in Equation 15, the average amount per transaction (AAPT) feature is the ratio of the total amount of transactions processed by a merchant to the total number of transactions. This feature has been added to the dataset in order to further investigate the relationship between the average monetary value of a merchant's transactions and default behavior.

$$AAPT = \frac{Amount}{Numberofairtimepins + Numberofloandisbursement + Numberofmerchantdeposits + Numberofloanrepayments} \tag{15}$$

3.3.2 Credit history

The credit history feature represents a merchant's proportion of paid loans to total loans taken out on the platform. This feature is added to the dataset to represent the payment history of the merchant's previous loans.

$$credithistory = \frac{Numberofpaidloans}{Numberofpaidloans + Numberofdelinquentloans + Numberofdefaultloans} \tag{16}$$

3.4 Exploratory data analysis

The objective of the Exploratory Data Analysis (EDA) chapter is to gain a comprehensive understanding of the dataset by conducting several analyses. These include examining the correlation between features, assessing feature importance, analyzing the default status of merchants, and investigating the distribution of the target outcome. Through these analyses, we aim to identify the most relevant features and the relationships between them in predicting the target outcome, which is the default status of a merchant.

3.4.1 Correlation plot

Figure 5 depicts the correlation of the training set's features, including the target variable "default status". The plot helps to better understand the relationship between the features and the target variable and to address collinearity. The correlation magnitude values described in Figure 6's legend range from 1 to -1. A magnitude of one indicates a strong positive linear relationship, a magnitude of one indicates a strong negative relationship, and a magnitude of zero indicates that there is no linear relationship between the two features. Because all of the features were numerical, the Pearson method was used to determine feature correlation.

Based on the correlation plot, the target variable has a high positive relationship with the following variables: Number of delinquent loans, Number of defaulted loans, and Average Amount per Transaction. On the other hand, it has a negative correlation to the following variables: Credit History and Number of Airtime Pins. These correlations are important to consider when building a credit scoring model as they provide insight into which variables are most strongly associated with default status.

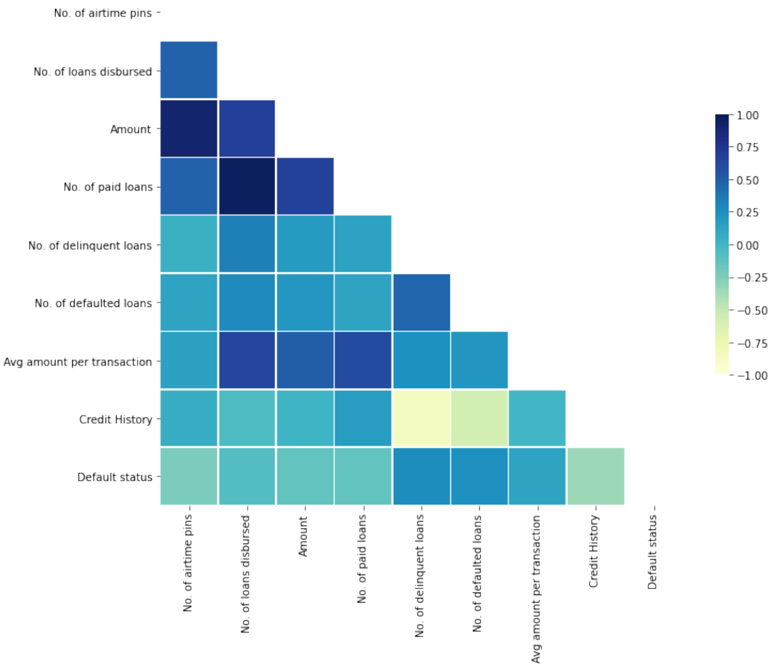

Figure 5: A correlation plot of the features

3.4.2 Feature importance

To determine the importance of the features, the random forests algorithm was used. The algorithm works by decorrelating the trees by selecting random predictors from the seven (7) features to split on for each tree and then predicting the target outcome for 1000 trees (number of estimators). Following that, the Gini importance of a feature is calculated as the decrease in node impurity weighted by the probability of reaching that node. The number of samples that reach the node divided by the total number of samples yields the node probability. The greater the value, the greater the importance of the feature in predicting the outcome [16].

According to Figure 6, the number of airtime pins, amount, and AAPT were the

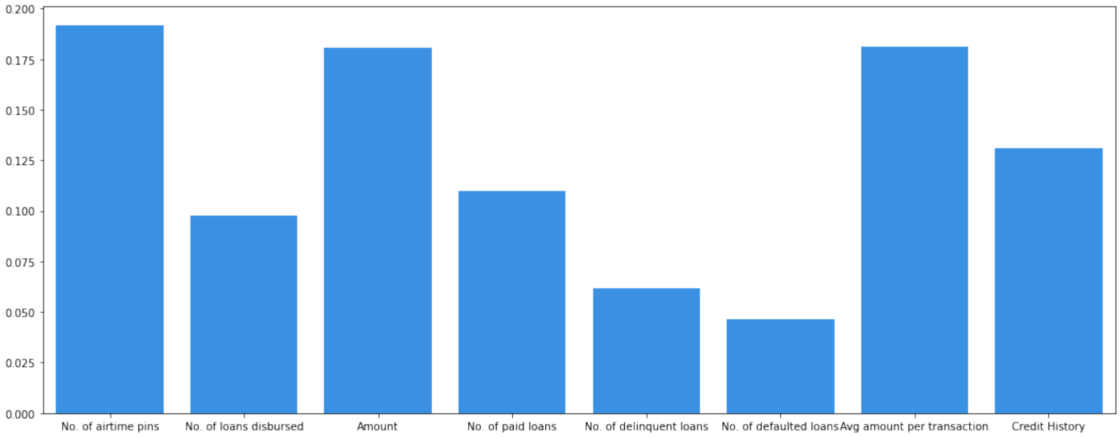

Figure 6: A bar plot showing feature Gini importance

most important, while the number of defaulted loans and number of delinquent loans were the least important.

3.4.3 Default status analysis

This section investigates how a merchant's default status relates to the engineered features discussed in Section 3.3 by plotting visualizations of the following relationships:

- Credit History and Default Status
- Credit History and Average Amount Per Transaction
- Default Status, Credit History and Average Amount Per Transaction

It is important to note that the observations made are based on a snapshot of the results for the period of April to November 2020. It is possible that the trends seen may vary if data from additional months were included in the analysis.

Credit history and default status - The credit history feature is a formula that identifies the percentage pay-back rate of a merchant's loans, and the default status feature describes the members who actually defaulted in November. The box plot of credit history versus default status in Figure 7 shows that:

- The interquartile range (IQR) for defaulters is 47%, while the IQR for non-defaulters is 20%. This suggests that defaulters have a credit history with a wider distribution than non-defaulters, indicating high variability.
- Merchants with a credit history greater than 80% are more likely to receive loans from lenders.
- Based on Figure 7, it appears that merchants with a credit history below 75% are more likely to default on their November loan. However, it should be noted that there is an outlier merchant with a 37% credit history who did not default on their November loans. It is possible that the trends may differ if data from additional months were included in the analysis.

Average amount per transaction (AAPT) and default status - The relationship between a merchant's transaction value and default behavior is demonstrated through a box plot of the average amount per transaction (AAPT) feature versus the default status feature, as shown in Figure 8. From the figure, it is observed that non-defaulters have an interquartile range (IQR) of 5.5, while defaulters have an IQR of 7, with a median of 16 and 18, respectively. The IQR for non-defaulters is smaller than the IQR of defaulters, indicating that the AAPT values for non-defaulters are more tightly clustered around the median value of 16 compared to defaulters whose AAPT values are more spread out around the median

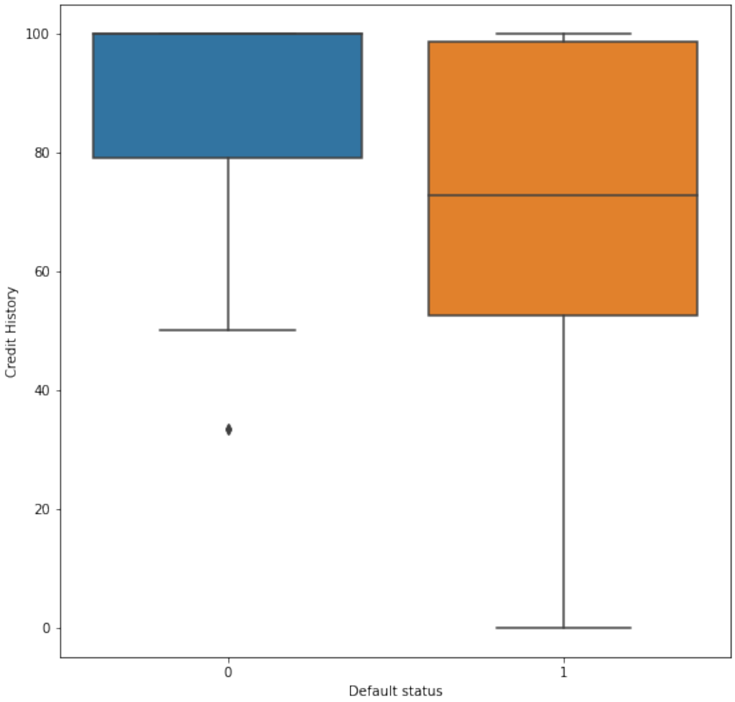

Figure 7: A box plot of Credit History feature against Default Status feature (non-defaulters in blue and defaulters in orange))

value of 18. Therefore, when a merchant's AAPT is greater than 18, they are more likely to default, but it is worth noting that this is not always the case, as there are still defaulters with lower AAPT values and non-defaulters with higher AAPT values.

This suggests that AAPT alone is not a strong predictor of default status and should be used in conjunction with other features for better prediction performance. Additionally, because the boxplot shows an overlap in distribution range for defaulters and non-defaulters, the AAPT feature alone is not a good indicator of merchant default status.

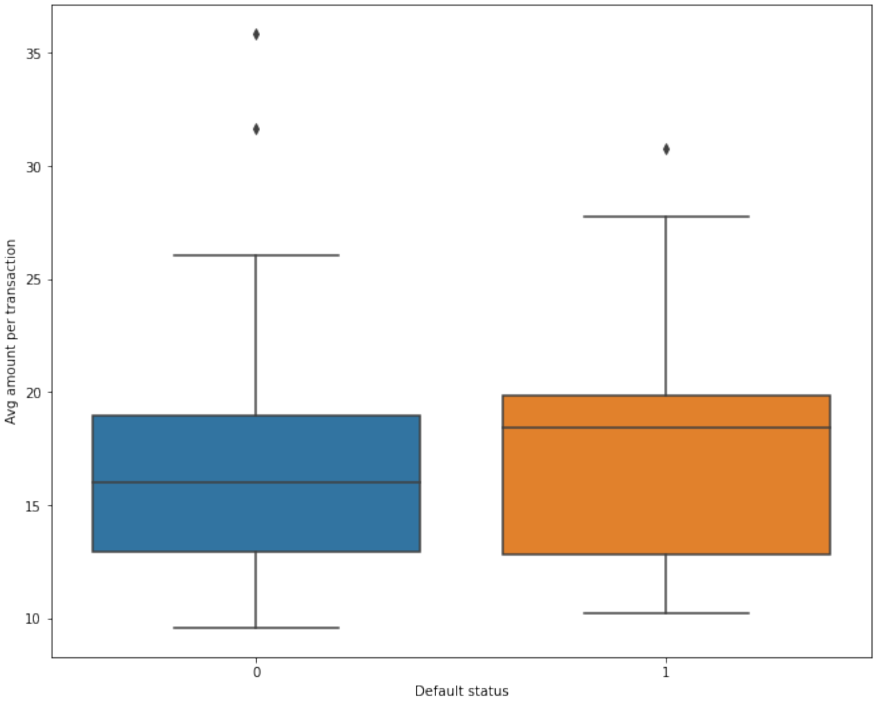

Figure 8: A box plots of Average Amount Per Transaction feature against Default Status (non-defaulters in blue and defaulters in orange)

Credit history, average amount per transaction (AAPT) and default status - Figure 9 combined the credit history, average amount per transaction and default status features in a scatter plot to investigate the relation between them. The following were the observations drawn:

- The defaulters are randomly distributed throughout the plot, demonstrating that there is no relationship between the credit history, average amount per transaction, and default status features. This observation supports what is seen in Figure 5.

- According to Section 3.4.3, a merchant with a credit history of less than 80 is more likely to default. However, as shown in Figure 9, merchants with an average amount per transaction of less than 15 and a credit history of less than 80 have a 50% default rate, compared to the 83% default rate for merchants with an average amount per transaction of more than 15.

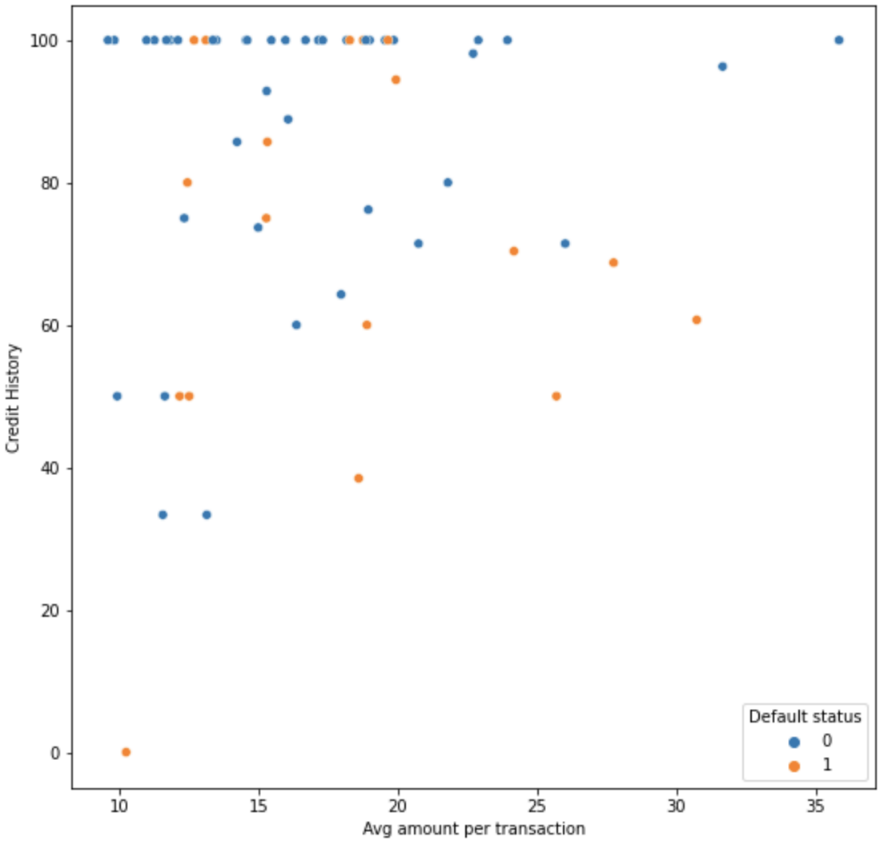

Figure 9: Scatter plot showing the distribution of non-defaulters (blue) and defaulters (orange) when Credit History feature is plotted against Average Amount Per Transaction feature

3.4.4 Class distribution of the target outcome

It is important to investigate for a class imbalance of the target outcome that can impact the model predictions.

The dataset utilized in this study consisted of a small sample size of only 62 observations, with a class distribution of 71% non-defaulters and 29% defaulters, as illustrated in Figure 10. It is important to note that this class imbalance has the potential to influence the model's predictive performance, as the model may be inclined towards predicting non-defaulters, resulting in reduced accuracy for defaulters. To address this issue, alternative evaluation metrics such as precision, recall, and F1 score should be taken into account.

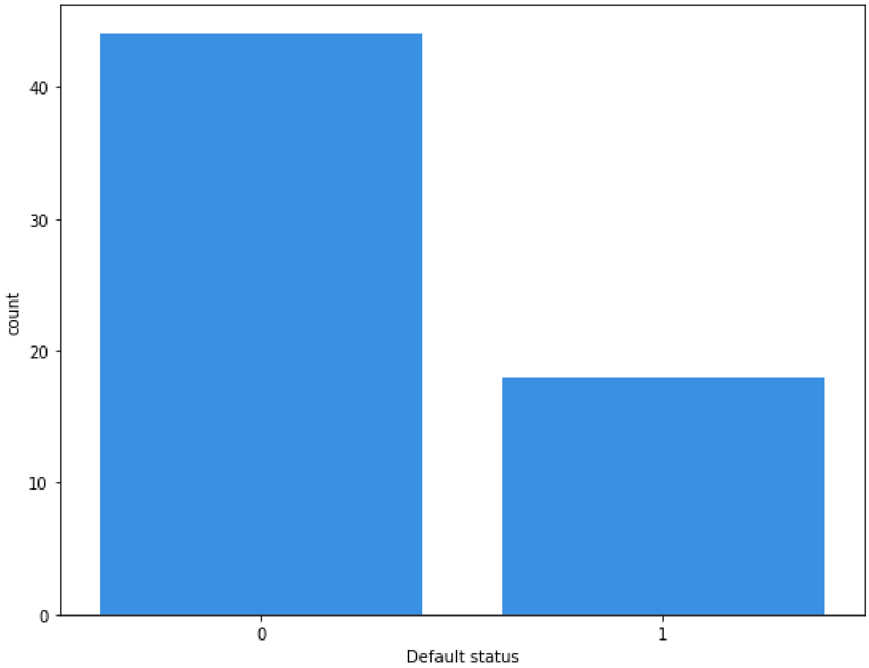

Figure 10: Bar plot showing the class distribution of the target outcome

3.5 Data pre-processing techniques

The following data preprocessing techniques were applied to the dataset:

- Number of merchant deposits and Number of loan repayments features have more than 80% missing values, and therefore were dropped for the modelling section.

- All of the numerical features were scaled to a standard normal distribution of mean of 0 and a standard deviation of 1.
- The data set was split into a 70%-30% train-test sub-samples.

3.6 Summary of methodology

The methodology section described how data from various sources was extracted, merged, and engineered into a single dataset used in Section 4. In addition, the section investigated and analyzed the data in order to fully understand its distribution and feature relationships. Finally, prior to the modelling section, preprocessing techniques were applied to the data.

In the next section, supervised learning methods such as logistic regression and support vector machines are applied on the data to predict the default status of the merchants.

4 Modelling

The modelling section of the research seeks to predict the outcome variable by using different sets of features. It consists of three main components: the feature combination selection process, modelling with logistic regression, and modelling with support vector machines (SVM). The feature selection process involves identifying the most appropriate feature combinations for the models.

4.1 Feature combination selection process

There are various statistical procedures used for selecting feature combinations that have a statistical significance with the target response [16]. The more flexible the model is, the more closely it resembles the data. In Section 3, the number of predictive features selected to be used for modelling were seven (7) which make a total of $2^7 or 128$ feature model combinations. As a result, based on the Akaike Information Criterion (AIC), the study selected the best five (5) models and added an extra all-feature combination - Combination 6 - as shown in Table 3.

Feature Combination Number	**Features**
Combination 1	Number of delinquent loans
Combination 2	Credit history
Combination 3	Average Amount Per Transaction & Credit history
Combination 4	Number of delinquent loans & Credit history
Combination 5	Number of paid loans & Numberof delinquent loans
Combination 6	All features

Table 3: Feature combination

From the model's loglikelihood and number of features, the AIC selects the best-fit model that explains the most variation with the fewest number of independent variables [16]. Please see Table 4, which show the model combinations based on AIC.

Selection Criterion	**Comb. 1**	**Comb. 2**	**Comb. 3**	**Comb. 4**	**Comb. 5**	**Comb. 6**
Degrees of freedom	2	2	3	3	3	8
Loglikelihood	-24.349	-24.528	-23.567	-23.643	-23.669	-22.317
AIC	52.997	53.335	53.749	53.902	53.952	61.833

Table 4: Feature combination selection criterion

4.2 Logistic regression

Logistic regression models the probability that the target response falls into one of two categories; default (1) or non-default (0). To fit the model to the dataset, a method called the maximum likelihood is used to estimate the β_p coefficients as shown in Table 5. The feature combination of the logistic regression models in Table 5 is in accordance with Table 3 such that the LR1 model uses feature Combination 1.

Table 5 shows the p-values in brackets for the independent variables with respect to the null hypothesis. The null hypothesis states that the independent variable (feature) and the dependent variable (target outcome) have no relationship [17]. As a result, a variable's level of statistical significance is expressed in p-values as follows:

- A p-value of less than 0.05 is statistically significant because it demonstrates strong evidence against the null hypothesis because the probability of the null hypothesis being correct is less than 5%. As a result, for all features/variables with a p-value less than 0.05, we reject the null hypothesis.
- A p-value greater than 0.05 is not statistically significant, since it demonstrates a strong evidence for the null hypothesis. Therefore, we accept the null

Model Number Features	LR1	LR2	LR3	LR4	LR5	LR6
(Intercept)	-0.9067 (0.0101)	-0.9414 (0.0082)	-0.9537 (0.0097)	-0.9781 (0.0081)	-1.0470 (0.0101)	1.0241 (0.0153)
Amount						-1.1392 (0.5740)
AAPT			0.4972 (0.0456)			1.2453 (0.2363)
Credit history		-0.5877 (0.0667)	-0.6519 (0.1732)	-0.4071 (0.2304)		-0.4998 (0.2776)
Number of airtime pins						0.7678 (0.6126)
Number of delinquent loans	0.6640 (0.0578)			0.4888 (0.1929)	0.8294 (0.0439)	0.5637 (0.4904)
Number of loan disbursements						-1.5011 (0.7273)
Number of paid loans					-0.71210 (0.3239)	0.7072 (0.8721)

Table 5: Shows the logistic regression models β_p coefficients with their p-values in brackets

hypothesis for all features/ variables with a p-value greater than 0.05.

The following example description can be used to interpret the logistic regression model coefficients β_p in Table 5;

- Increasing the number of delinquent loans by one unit while keeping all variables constant raises the probability of default by a multiple of 0.6640 in the LR1 model. As a result, the more the number of delinquent loans a merchant has, the greater the likelihood of default.
- For the LR3 model the following applies;
  - The probability of default rise by a multiple of 0.4972 when the Average Amount Per Transaction (AAPT) feature is increased by one unit and all predictors remain constant.

- When the Credit history feature is increased by one unit and all other predictors remain constant, the odds of default are reduced by a factor of 0.6519.

4.3 Support vector machines

Support vector machines (SVMs) seek out separators with the greatest margin in order to improve the classifier's generalization performance; as a result, they are also known as a large margin classifier because they create the greatest possible distance in the examples in the classification problem [16]. They use support vectors to discover the hyperplane while correctly classifying as many training points as possible.

The support vectors are observations that are either on the separator margin or on the wrong side of the separator but still within the maximum margin. The C (Cost) hyper-parameter controls the number of support vectors. Misclassification of data on the incorrect side of the margin separator is penalized with a high C (Cost), resulting in low bias and high variance [22].

Also, even if the original data are non-separable, SVM kernel methods effectively transform the input data into a high-dimensional space where a linear separator may exist. Data that are not linearly separable in the original input space are often easily separable in a higher-dimensional space. In the original space, the high-dimensional linear separator is actually nonlinear. Kernels are a function that computes the inner products in transformed space and quantify the similarity between observations [10]. The study used the linear and radial based function (RBF) kernels in a trial and error method, by fitting the kernels to all the feature combinations. The RBF kernel has an extra hyperparameter called γ (Gamma) because it is a non-linear separator. The higher the γ (Gamma) value is the more the SVM tries to fit the training data. Thus, a high γ (Gamma) results in high variance and low bias [22].

The grid-search method was used to generate the C (Cost) and γ (Gamma) values in Table 6, with C (Cost) ranging from 0.1 to 100 in 0.5 steps and γ (Gamma) values of 0.1, 1, 10 and 100 were used.

Model Number	**Kernel**	**C, Cost**	**γ, Gamma**	**Number of vectors**
SVM 1A	Linear	0.1	-	26
SVM 1B	Radial	1.1	10	27
SVM 2A	Linear	0.1	-	27
SVM 2B	Radial	27.1	1	27
SVM 3A	Linear	0.1	-	27
SVM 3B	Radial	1.1	10	34
SVM 4A	Linear	0.1	-	27
SVM 4B	Radial	17.6	0.1	23
SVM 5A	Linear	0.1	-	28
SVM 5B	Radial	5.6	0.1	28
SVM 6A	Linear	79.1	-	27
SVM 6B	Radial	2.6	0.1	31

Table 6: SVM model hyperparameters summary table

4.4 Summary of modelling

The modelling section used a feature selection method based on the Akaike Information Criterion (AIC) on seven (7) features to generate five (5) models from 128 different feature combinations that have the least complexity but explain the most variation in the data. In addition, an additional all-feature model was added to the sixth feature combination. Following that, modelling was carried out using the logistic regression and support vector machines methods. The modelling results of this section will be reported in the following section.

5 Results

The models built in Section 4 will be discussed in this section, and their performance will be evaluated using the following metrics:

- Train Accuracy - The model's prediction accuracy on the training data.
- Leave-One-Out-Cross-Validation (LOOCV) - The prediction accuracy of a model in which each observation is regarded a validation set and the rest is a training set ($N - 1$, where N is the number of observations). Because the data set is small (only 62 observations), this method was chosen over the well-known k-fold cross-validation method.
- Test Accuracy - The model's accuracy in predicting the outcome of the randomly chosen 25% of the dataset, which was used as the testing set.
- Gini coefficient - A linear transformation of the area under the curve (AUC) of a receiving operating characteristic (ROC) curve as explained in Section 2.2.3, is a good metric for classification models.

5.1 Logistic regression

Tables 10 through 27 in Section 8 provide a full description of the Table 7 prediction results.

It should be noted that the inconsistency in the evaluation metrics of the logistic regression models presented in Table 7 is due to the small sample size and imbalanced data set. The LR6 model, which has the highest test accuracy but poor cross-validation accuracy, and the LR3 model, which has high training, cross-validation, and test accuracy but a low Gini coefficient, demonstrate the limitations in predicting unseen data.

On the other hand, the LR2 model, which outperformed the others in cross-validation accuracy and Gini coefficient metrics, is the best model for data generalization. This is confirmed by the feature selection approach described in Section 4.1 and illustrated in Table 4, where LR1 and LR2 have the lowest AIC values. Nonetheless, the imbalanced and small sample size of the data set may have influenced the models' performance, indicating the need for caution when interpreting the results.

Model Number	**Train Accuracy**	**LOOCV Accuracy**	**Test Accuracy**	**Gini Coefficient**
LR1	0.7674	0.6935	0.6316	0.4857
LR2	0.6977	0.7258	0.7368	0.6143
LR3	0.7674	0.7096	0.7368	0.3142
LR4	0.7907	0.6290	0.6316	0.4571
LR5	0.7442	0.6774	0.6316	0.3571
LR6	0.7674	0.6774	0.7895	0.4285

Table 7: Logistic regression models evaluation metrics

5.2 Support vector machines

Tables 28 to Table 52 in Section 8 provide a detailed description of the prediction results for the support vector machines models in Table 8. However, it should be noted that there is inconsistency in the evaluation metrics, as seen from the confusion matrices, with only three (3) models (SVM 2B, SVM 6A, and SVM 6B) predicting true positives out of the twelve (12) SVM models evaluated.

Based on the test accuracy and Gini coefficient metrics, it is observed that models with linear kernels performed better than those with radial kernels. However, some models with high cross-validation accuracies, such as SVM 4B and SVM 1B, showed low test accuracy and average Gini coefficients, indicating inconsistency in their performance on the evaluation metrics. The inconsistency in the model evaluation metrics may be due to the small sample size of the dataset.

Nonetheless, the all-feature combination model SVM 6A with a linear kernel, with a Gini coefficient of 0.6000, demonstrated the best performance among the SVM models.

Model Number	**Train Accuracy**	**LOOCV Accuracy**	**Test Accuracy**	**Gini Coefficient**
SVM 1A	0.6977	0.7097	0.7368	0.4857
SVM 1B	0.7907	0.7258	0.6842	0.4857
SVM 2A	0.6977	0.6774	0.7368	0.4857
SVM 2B	0.7674	0.6935	0.7368	0.4857
SVM 3A	0.6977	0.6774	0.7368	0.4857
SVM 3B	0.8837	0.6290	0.6316	0.2285
SVM 4A	0.6977	0.7097	0.7368	0.4857
SVM 4B	0.7907	0.7258	0.6316	0.3428
SVM 5A	0.6977	0.7097	0.7368	0.4857
SVM 5B	0.7209	0.6613	0.6316	0.3428
SVM 6A	0.76747	0.6290	0.7895	0.6000
SVM 6B	0.8140	0.7097	0.6842	0.4857

Table 8: SVM models' train and test dataset prediction accuracy

5.3 Summary of the results

This section presents a summary of the logistic regression and support vector machines models evaluated in Section 4 using four evaluation metrics: train accuracy, leave-one-out cross-validation (LOOCV) accuracy, test accuracy, and Gini coefficient. The support vector machines models outperformed the logistic regression models, with an average Gini coefficient of 0.4499 compared to the logistic regression models' average of 0.4427, as calculated from Table 7 and Table 8 .

However, it should be noted that there is inconsistency in the evaluation metrics of the logistic regression models in Table 7 and the SVM models in Table 8 due to the small sample size and an imbalanced data set. For example, the LR6 model has the highest test accuracy but poor cross-validation accuracy, while the LR3 model has high training, cross-validation, and test accuracy but a low Gini coefficient, indicating limitations in predicting unseen data.

According to the findings of this study, the logistic regression model LR2 demonstrated superior performance in comparison to the other models. The LR2 model exhibited a leave-one-out cross-validation accuracy of 0.7258 and a Gini coefficient of 0.6143. These results were further supported by the feature selection approach described in Section 4.1 and presented in Table 4. Specifically, LR1 and LR2 had the lowest AIC values, suggesting that the LR2 model is the most suitable for generalizing the data. However, it is worth noting that the small sample size and class imbalance of the data set may have impacted the models' performance, indicating that the results should be interpreted with caution.

The following section will provide an in-depth analysis of the LR2 model's performance in predicting default and assessing creditworthiness for informal merchants.

6 Credit score model analysis

According to the study's analysis and results, LR2 is the best performing model. As described in Section 4.2, this model has only one feature: the merchant's credit history, which is used to predict the merchant's likelihood of default. Table 9 depicts the distribution of merchants in terms of the probability of default, and it shows that when the probability of default is greater than 50%, the default rate is on average at 78%, while when the probability of default is less than 50%, the default rate is on average at 21%.

Predicted probability of default	Defaulters	Number of merchants	Default rate
0% - 10%	0	7	0%
11% - 20%	4	18	22%
21% - 30%	3	13	23%
31% - 40%	3	8	38%
41% - 50%	1	7	14%
51% - 60%	3	4	75%
61% - 70%	0	0	Unknown
71% - 80%	2	3	67%
81% - 90%	2	2	100%
91% - 100%	0	0	Unknown

Table 9: Showing the predicted probability of defaulters of the merchants as per the selected credit scoring model for a loan decision

Figure 11 depicts a bar plot distribution of defaulters and non-defaulters based on the observations in Table 9. The distribution of defaulters and non-defaulters within the probability of default ranges contributes to the Table 9 results and identifies the credit decision to be made for a given probability of default. As a result of Table 9 and Figure 11, the following findings were obtained:

- Default probability ranges of 0% - 10% provide strong evidence of approval of all loan applications in that range.

- The probability of default ranges of 11% - 30% provide marginally good evidence to approve loan applications in that range, because they have a combined average default rate of 22.5% which is derived from Table 9.
- The probability of default ranges of 31%-40% and 41%-50% show an irregularity in the default rates, as the 41%-50% range was expected to have a higher default rate than the 31%-40% range. As a result, for merchants who fall within the given probability range, it is recommended that the lender perform an additional assessment prior to credit advancement.
- Because the number of defaulters exceeds the number of non-defaulters, probability of default ranges of 51% - 100% provide a strong evidence for rejecting all loan applications that fall within that range.

It is important to note that the credit score model's performance and results may improve with more data. However, even with the small dataset used in this study, there is a clear shift in the proportion of defaulters to non-defaulters across the 50% probability of default mark. This indicates the potential usefulness of the credit score model as a decision-making tool for assessing the creditworthiness of informal merchants.

As a result, the decision-making tool presented in Figure 12 serves as an outcome for real-time credit decisions at scale, aiming to minimize credit risk and ensure effective management. It is derived from the results of the best performing model, LR2, as shown in Figure 11. To enhance its reliability, future studies should validate the tool's performance with a larger and balanced dataset, thereby optimizing its effectiveness in making informed credit decisions while mitigating associated risks.

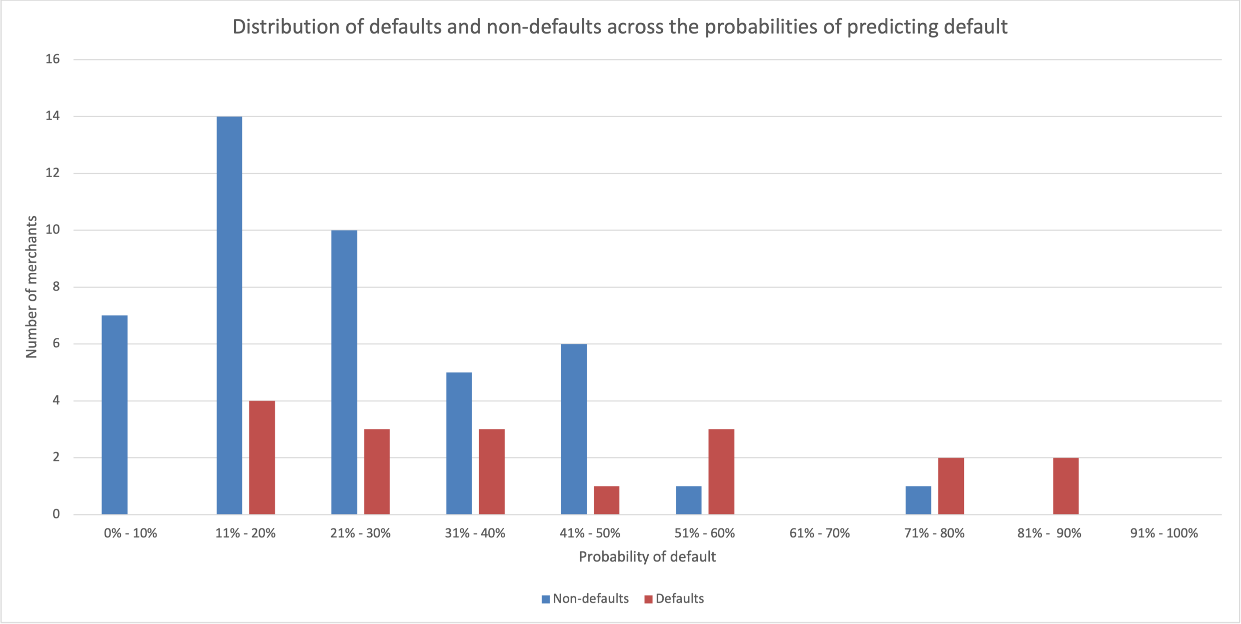

Figure 11: A bar plot showing the distribution of defaulters and non-defaulters as credit scoring model

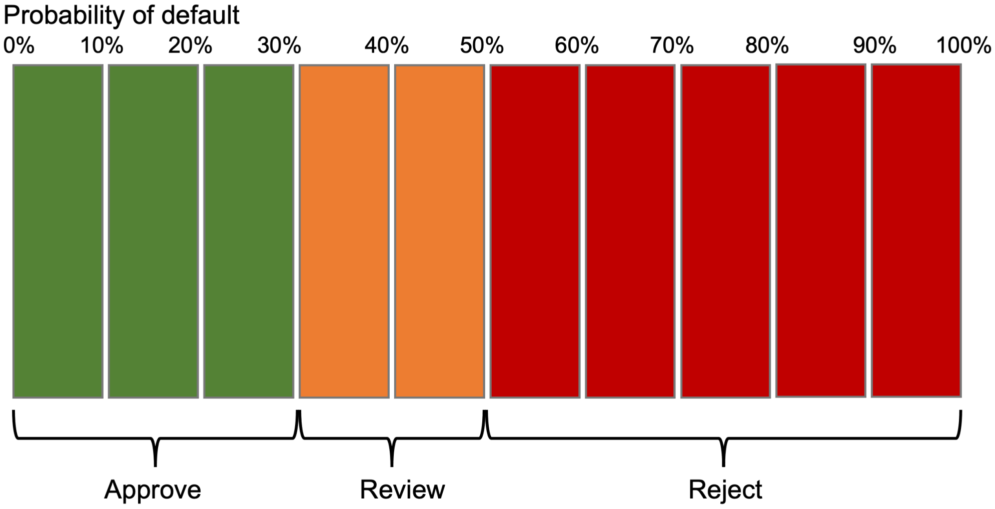

Figure 12: A figure showing the credit decision-making tool from the probability of default of the credit scoring model

Note: The heights of the bars in this diagram are provided as a visual representation and do not hold direct significance in terms of variations or specific values.

7 Conclusion

The study developed credit scoring models to predict default by using transactional and loan data from informal merchants. Seven (7) predictive features were extracted and processed, and an AIC feature selection method was used to obtain five feature combinations for modelling and analysis, including a sixth feature combination with all seven (7) features.

The study evaluated six (6) logistic regression models and twelve (12) support vector machine models for their default predictive power on informal merchants. Support vector machines outperformed logistic regression models based on the average performance on the Gini coefficient metric. However, the best-performing model was a logistic regression model with a merchant's credit history as the only feature, which resulted in a 0.6143 Gini coefficient.

Considering the limited dataset of 62 informal merchants' transactional and loan data used in this study, it is important to note that the best-performing model was observed to be the one that incorporated the merchant's credit history. While this finding highlights the significance of credit history as a determinant of default, it is essential to exercise caution when generalizing these results due to the small sample size. The implications for the generalizability of the findings should be taken into consideration, and further research with larger and more diverse datasets would be valuable to validate and extend these observations.

However, the methodology used in this study, which includes feature selection and modelling with logistic regression and support vector machines, can be replicated with continued collection of new data from Nomanini's operations to create a more robust credit scorecard for informal merchants. Such an approach can provide financial institutions with valuable insights when extending credit and managing risk to this important segment of the economy.

7.1 Limitations of study

This study, like the majority of studies, has limitations. Only 62 merchants took out loans in November 2020, yielding a dataset of 62 observations. As a result, only a few modelling techniques, such as logistic regression and support vector machines, were used in the study, while others, such as decision trees, random forests, ensemble methods, and neural networks, were omitted. The use of more advanced modelling techniques would allow a comparison evaluation based on criteria such as prediction power and model complexity.

7.2 Areas of further research

This study's research addresses the research objectives stated in Section 1.3. However, there are a number of areas that can be further investigated in order to develop more robust credit scoring models, such as the following:

- The study used a small dataset of 62 observations (merchants), which led to the failure of conducting modelling and analyses with advanced machine learning methods such as decision trees, random forests, ensemble methods, and neural networks.
- This study focused on merchant behavioral scoring rather than application scoring. As a result, more research on features to be used when a merchant first applies for a loan should be conducted.

- Implementation of survival modelling analysis on data to address recurring merchant default or delinquency events.

References

[1] Hussein A. Abdou and John Pointon. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3):59–88, 2011.

[2] Bart Baesens. *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. John Wiley Sons, 2014.

[3] Bart Baesens, Daniel Roesch, and Harald Scheule. *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. John Wiley Sons, 2016.

[4] Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, 2003.

[5] Concepcio Bartual Sanfeliu, Fernando Garcia Garcia, Francisco Guijarro, and Agustin Romero Civera. Probability of default using the logit model: the impact of explanatory variable and database selection. In *International Scientific Conference: Whither Our Economics*, pages 118–124. Mykolas Romeris University, 2012.

[6] Joao Bastos. Credit scoring with boosted decision trees. *Journal of the Operational Research Society*, 58(8):1059–1067, 2007.

[7] Amrei Botha and Devon Natasha Maylie. The msme voice: Growing south africa's small business sector. Technical report, The World Bank, 2020.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] Konstantinos G. Derpanis. Mean shift clustering. *Lecture Notes*, page 32, 2005.

[10] Jerome H Friedman. *The elements of statistical learning: Data mining, inference, and prediction.* springer open, 2017.

[11] Rui Ying Goh and Lai Soon Lee. Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, 2019, 2019.

[12] Siana Halim and Yuliana Vina Humira. Credit scoring modeling. *Jurnal Teknik Industri*, 16(1):17–24, 2014.

[13] David J Hand. Classifier technology and the illusion of progress. *Statistical Science*, pages 1–14, 2006.

[14] David J. Hand and William E. Henley. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.

[15] David J. Hand and Saul D. Jacka. *Statistics in Finance.* Chapman Hall, 1998.

[16] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

[17] David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3), 2007.

[18] Leandro Medina, Aleksandr W. Jonelis, and Merve Cangul. *The Informal Economy in Sub-Saharan Africa: Size and Determinants.* International Monetary Fund, 2017.

[19] Nomanini. Fintech the informal retail economy in africa, 2019.

[20] Nomanini. Supply chain financing for informal msmes, 2020.

[21] James A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, pages 109–131, 1980.

[22] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jacob Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Stuart Russell and Peter Norvig. Ai a modern approach. *Learning*, 2(3):4, 2005.

[24] Mark Schreiner. Scoring: the next breakthrough in microcredit. *Occasional Paper*, 7, 2003.

[25] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[26] Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Springer, 2008.

[27] Lyn C. Thomas. *Credit Scoring and its Applications.* SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000.

[28] Maria Fernandez Vidal and Fernando Barbon. Credit scoring in financial inclusion. *Journal of Financial Services Marketing*, 24(3):145–156, 2019.

[29] John C. Wiginton. A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, pages 757–770, 1980.

[30] Qiong Zhang. Modeling the probability of mortgage default via logistic regression and survival analysis. 2015.

8 Appendix

8.1 Github repository

Please find the datasets and the notebooks in the following github link - `https://github.com/kazimoto11/fintech_thesis`.

8.2 Logistic regression tables

Feature	**Estimate**	**Standard Error**	**Z-Value**	**Pr**$(>|z|)$
(Intercept)	-0.9067	0.3525	-2.572	0.0101
Number of delinquent loans	0.6640	0.3500	1.897	0.0578

Table 10: Model LR1 training set logistic regression coefficients summary

	Predicted 0	**Predicted 1**
Actual 0	29	1
Actual 1	9	4

Table 11: Model LR1 training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	12	2
Actual 1	5	0

Table 12: Model LR1 testing set confusion matrix

Feature	**Estimate**	**Standard Error**	**Z-Value**	**Pr**$(>|z|)$
(Intercept)	-0.9414	0.3559	-2.645	0.0082
Credit history	-0.5877	0.3205	-1.834	0.0667

Table 13: Model LR2 training set logistic regression coefficients summary

	Predicted 0	**Predicted 1**
Actual 0	28	2
Actual 1	11	2

Table 14: Model LR2 training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	5	0

Table 15: Model LR2 testing set confusion matrix

Feature	**Estimate**	**Standard Error**	**Z-Value**	**Pr**$(>\|z\|)$
(Intercept)	-0.9537	0.3689	-2.585	0.0097
Credit history	-0.6519	0.3260	-1.999	0.0456
AAPT	0.4972	0.3650	1.362	0.1732

Table 16: Model LR3 training set logistic regression coefficients summary

	Predicted 0	**Predicted 1**
Actual 0	28	2
Actual 1	8	5

Table 17: Model LR3 training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	13	1
Actual 1	4	1

Table 18: Model LR3 testing set confusion matrix

Feature	**Estimate**	**Standard Error**	**Z-Value**	**Pr**$(>\|z\|)$
(Intercept)	-0.9781	0.3694	-2.648	0.0081
Credit history	-0.4071	0.3395	-1.199	0.2304
Number of delinquent loans	0.4888	0.3755	1.302	0.1929

Table 19: Model LR4 training set logistic regression coefficients summary

	Predicted 0	**Predicted 1**
Actual 0	29	1
Actual 1	8	5

Table 20: Model LR4 training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	12	2
Actual 1	5	0

Table 21: Model LR4 testing set confusion matrix

Feature	**Estimate**	**Standard Error**	**Z-Value**	**Pr**$(>\|z\|)$
(Intercept)	-1.0474	0.4074	-2.571	0.0101
Number of paid loans	-0.7121	0.7219	-0.986	0.3239
Number of delinquent loans	0.8294	0.4116	2.015	0.0439

Table 22: Model LR5 training set logistic regression coefficients summary

	Predicted 0	**Predicted 1**
Actual 0	29	1
Actual 1	10	3

Table 23: Model LR5 training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	12	2
Actual 1	5	0

Table 24: Model LR5 testing set confusion matrix

Feature	**Estimate**	**Standard Error**	**Z-Value**	**Pr**$(>\|z\|)$
(Intercept)	-1.0241	0.4222	-2.426	0.0153
Number of airtime pins	0.7678	1.5162	0.506	0.6126
Number of loan disbursements	-1.5011	4.3044	-0.349	0.7273
Amount	-1.1392	2.0267	-0.562	0.5740
Number of paid loans	0.7072	4.3920	0.161	0.8721
Number of delinquent loans	0.5637	0.8174	0.690	0.4904
AAPT	1.2453	1.0515	1.184	0.2363
Credit history	-0.4998	0.4603	-1.086	0.2776

Table 25: Model LR6 training set logistic regression coefficients summary

	Predicted 0	**Predicted 1**
Actual 0	28	2
Actual 1	8	5

Table 26: Model LR6 training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	13	1
Actual 1	3	2

Table 27: Model LR6 testing set confusion matrix

8.3 Support vector machines tables

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	13	0

Table 28: Model SVM 1A training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	5	0

Table 29: Model SVM 1A testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	9	4

Table 30: Model SVM 1B training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	5	0

Table 31: Model SVM 1B testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	13	0

Table 32: Model SVM 2A training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	5	0

Table 33: Model SVM 2A testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	29	1
Actual 1	9	4

Table 34: Model SVM 2B training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	13	1
Actual 1	4	1

Table 35: Model SVM 2B testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	13	0

Table 36: Model SVM 3A training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	5	0

Table 37: Model SVM 3A testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	13	0

Table 38: Model SVM 3B training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	5	0

Table 39: Model SVM 3B testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	13	0

Table 40: Model SVM 4A training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	5	0

Table 41: Model SVM 4A testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	9	4

Table 42: Model SVM 4B training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	12	2
Actual 1	5	0

Table 43: Model SVM 4B testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	13	0

Table 44: Model SVM 5A training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	5	0

Table 45: Model SVM 5A testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	29	1
Actual 1	11	2

Table 46: Model SVM 5B training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	12	2
Actual 1	5	0

Table 47: Model SVM 5B testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	29	1
Actual 1	9	4

Table 48: Model SVM 6A training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	4	1

Table 49: Model SVM 6A testing set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	30	0
Actual 1	11	2

Table 50: Model SVM 6B training set confusion matrix

	Predicted 0	**Predicted 1**
Actual 0	14	0
Actual 1	4	1

Table 51: Model SVM 6B testing set confusion matrix

8.4 Merchant probability of default table

S/N	Merchant Account Id	Probability of Default
1	05f2741a39f5446eb0068f18672854aa	48%
2	065d13d0cfbf4df4b99258ffa25f9a65	20%
3	0772a5e473344620b882907c518d5165	20%
4	0a8afd3da40a4d39a44c919f1a7753d4	77%
5	0e7e5806161844ba8105a9b26f467475	30%
6	17315aa236a5400489347b99298d6252	25%
7	1cd06eb5b0294306a3e081e744cc0f25	27%
8	20644c99212744e8b3582a9d804afc87	20%
9	20dbd8740dd74ff687c0a171a3762702	20%
10	264a5c4a06e243578fa51f7ec69ca283	24%
11	2c617cf6a2f64c7e8a252f7f4dc9e13b	33%
12	2e5577f0bb10404dac0acdd0b1470d3d	35%
13	3035b8c6e86445dc8bc788d4dfae5d00	42%
14	3511ee5918e1438e908239f1f12b0267	20%
15	36418871a5d84de8ba7e31fe02f9f375	23%
16	397454beef294ce697dfdbbdde0649bb	20%
17	3ef69167961c425c852bac73aab030f2	20%
18	3f556e6e42af4393ad2c3676fec34d7a	35%
19	42ea5ec094784f2fba5bf91b26909c49	30%
20	49587605d78e4e6ebf0a98477cc2de34	48%
21	4d4618e743484e43a8391b3f87bdf060	20%
22	500a70aed605442886a25563e7624540	48%
23	55e47faa2eeb43569f45f51d0ecc5f51	33%
24	5911724549704f858d8ecbe1606ac613	48%
25	5ccbb392cc3741fd992c37543c0f1152	20%
26	5f5e94a3123d40b28ce87569c729d10e	59%
27	5f8927cf5c3c4a6abdd82949906b0440	20%
28	61c15b22d6f9425ea3a64d7b6cc96e52	20%
29	65089c9b6bc34baa81937446e0d539f2	20%
30	7365f4f8b4e04195bdac220d9d201e44	20%
31	743e6b791db949a29a2f7ed5e78bc155	20%

S/N	Merchant Account Id	Probability of Default
32	750264e2a9b24671ad9e3e33086864d8	20%
33	77982336cebf4286b30f944ee5a785f8	39%
34	78f1d5858ffd4f968b45b4baee977907	20%
35	7b0d6876237e4b23abd21bcb2ffbaa11	20%
36	7c2cd2356226499d89082ca218dead90	20%
37	7f521dd99d0d4b64a0013f31ea25801e	21%
38	81da666bd316479487f2b7fae287e36d	20%
39	83586bc3d6244f09aa4ec25e651bfccc	20%
40	8c31a661eb4648ed898fc229519b2e4b	20%
41	8c4fe92d808a44d7a040c391414e41c9	41%
42	8cb06e61bdd740719718075e7f5a7744	20%
43	9b3c41aad63644d184bb0ab6a03fb6c4	59%
44	9ee35356e7c545b28edfd44ed81af3ac	36%
45	a7d997d93a714fa2a7ec9388c9fc2cd4	20%
46	b14349d48a084e81b3ec355eafeb5a41	35%
47	b8140e745b07457699c128b075a16ee4	27%
48	bca3a5bc2a3846669b3bf54e8a72bc53	20%
49	bde0f61714b24dcaa2fc21a3204229ba	22%
50	c0506ebcc4be424487a515e251c96927	33%
51	c6a5afe385b04dc4884b740089f4beaa	20%
52	cf583e18df804089ae2666d6d70d3570	20%
53	d2a7a55400dd4233b4b8d3ab643dc76c	55%
54	d61898f95f274288873ab1c4b4146717	20%
55	d812384df75c4e2e8b8e469fb6d0fd59	32%
56	dbc49cc6325240e3a87cea7d57ff1d14	20%
57	de9280bcfa834d77bfa1c658e900fa0e	42%
58	e10b6359478143ba98ec19c063e8fbc3	20%
59	e331e804730749c388effcc58d5bd04b	20%
60	e5fc362b25f945eabb9c5b6aa4362749	20%
61	f20cdccd212d4410a02237eb325a2ff5	48%
62	fd5bcdd080c5441caccef843042dfc6e	20%

Table 52: Showing the probability of default of merchants