



UNIVERSITY OF CAPE TOWN

ECO 5016W

MINOR DISSERTATION IN FINANCIAL TECHNOLOGY

Using Supervised Learning Methods to Credit Score Informal Merchants

Student:

Derick Kazimoto

Supervisor:

Dr. Şebnem Er

Student Number:

KZMDER001

Co-Supervisor:

A/Prof. Co-Pierre Georg

February 4, 2022

Contents

1	Introduction	13
1.1	Research Motivation	13
1.2	Background of Study	13
1.3	Objectives of Research	14
1.4	Limitations of Study	14
2	Literature review	15
2.1	Credit scoring in informal markets	15
2.1.1	Nomanini	15
2.1.2	Financial technology and the informal retail economy in Africa	18
2.2	Building credit scoring models	19
2.2.1	Data sources and quality	19
2.2.2	Credit scoring models	21
2.2.3	Evaluating a credit score model	28
2.3	Summary of literature	33
3	Methodology	35
3.1	Data extraction	35
3.1.1	Transaction data	35
3.1.2	Loan data	36
3.2	Merging the datasets	40
3.3	Feature engineering	41
3.3.1	Average Amount Per Transaction (AAPT) feature	41

3.3.2	Credit history	41
3.4	Exploratory data analysis	42
3.4.1	Correlation plot	42
3.4.2	Feature importance	43
3.4.3	Default status analysis	44
3.4.4	Class distribution of the target outcome	48
3.5	Data pre-processing techniques	48
3.6	Summary of methodology	49
4	Modelling	50
4.1	Feature selection	50
4.2	Logistic regression	51
4.3	Support vector machines	53
4.4	Summary of modelling	54
5	Results	55
5.1	Logistic regression	55
5.2	Support vector machines	56
5.3	Summary of the results	57
6	Credit score model analysis	59
7	Conclusion	62
7.1	Limitations of study	62
7.2	Areas of further research	63
8	Appendix	68

8.1	Logistic regression tables	68
8.2	Support vector machines tables	70
8.3	Merchant probability of default table	74

List of Tables

1	Comparison between judgemental scoring and statistical scoring [25].	19
2	Confusion Matrix	29
3	Feature combination	50
4	Feature combination selection criterion	50
5	Shows the logistic regression models β_p coefficients with their p-values in brackets	51
6	SVM model hyperparameters summary table	54
7	Logistic regression models evaluation metrics	56
8	SVM models' train and test dataset prediction accuracy	57
9	Showing the predicted probability of defaults of the merchants as per the the selected credit scoring model for a loan decision	59
10	Model LR1 training set logistic regression coefficients summary	68
11	Model LR1 training set confusion matrix	68
12	Model LR1 testing set confusion matrix	68
13	Model LR2 training set logistic regression coefficients summary	68
14	Model LR2 training set confusion matrix	68
15	Model LR2 testing set confusion matrix	68
16	Model LR3 training set logistic regression coefficients summary	69
17	Model LR3 training set confusion matrix	69
18	Model LR3 testing set confusion matrix	69
19	Model LR4 training set logistic regression coefficients summary	69
20	Model LR4 training set confusion matrix	69
21	Model LR4 testing set confusion matrix	69

22	Model LR5 training set logistic regression coefficients summary	69
23	Model LR5 training set confusion matrix	70
24	Model LR5 testing set confusion matrix	70
25	Model LR6 training set logistic regression coefficients summary	70
26	Model LR6 training set confusion matrix	70
27	Model LR6 testing set confusion matrix	70
28	Model SVM 1A training set confusion matrix	70
29	Model SVM 1A testing set confusion matrix	71
30	Model SVM 1B training set confusion matrix	71
31	Model SVM 1B testing set confusion matrix	71
32	Model SVM 2A training set confusion matrix	71
33	Model SVM 2A testing set confusion matrix	71
34	Model SVM 2B training set confusion matrix	71
35	Model SVM 2B testing set confusion matrix	71
36	Model SVM 3A training set confusion matrix	71
37	Model SVM 3A testing set confusion matrix	72
38	Model SVM 3B training set confusion matrix	72
39	Model SVM 3B testing set confusion matrix	72
40	Model SVM 4A training set confusion matrix	72
41	Model SVM 4A testing set confusion matrix	72
42	Model SVM 4B training set confusion matrix	72
43	Model SVM 4B testing set confusion matrix	72
44	Model SVM 5A training set confusion matrix	72
45	Model SVM 5A testing set confusion matrix	73
46	Model SVM 5B training set confusion matrix	73

47	Model SVM 5B testing set confusion matrix	73
48	Model SVM 6A training set confusion matrix	73
49	Model SVM 6A testing set confusion matrix	73
50	Model SVM 6B training set confusion matrix	73
51	Model SVM 6B testing set confusion matrix	73
52	Showing the probability of default of merchants	75

List of Figures

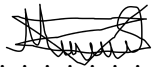
1	An example of a decision tree	24
2	Figure 2(a) shows two classes of points (black and white circles) and three candidate linear separators. Figure 2(b) shows the maximum margin separator (heavy line) is at the midpoint of the margin between dashed lines. The support vectors are the points with large circles on the dashed lines as the closest examples to the separator[24].	26
3	Figure 3(a) shows a two-dimensional training dataset of black and white circles with a decision boundary of $x_1^2 + x_2^2 \leq 1$. Figure 3(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 3(a) is transformed into a linear decision boundary in Figure 3(b) [24].	27
4	Example of a ROC curve [28].	32
5	Example of a Kolmogorov-Smimov curve [28].	34
6	A correlation plot of the features	42
7	A bar plot showing feature Gini importance	43
8	A box plot of Credit History feature against Default Status feature (non-defaulters in blue and defaulters in orange))	45
9	A box plots of Average Amount Per Transaction feature against Default Status (non-defaulters in blue and defaulters in orange)	46
10	Scatter plot showing the distribution of non-defaulters (blue) and defaulters (orange) when Credit History feature is plotted against Average Amount Per Transaction feature	47
11	Bar plot showing the class distribution of the target outcome	48

12	A bar plot showing the distribution of defaulters and non-defaulters as credit scoring model	60
13	A figure showing the credit decision analysis from the probability of default of the credit scoring model	61

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: **04/February/2022**

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 barrier to growth of the informal retail sector. Due to a financial mismatch between the credit products offered, informal merchants are unable to obtain credit from financial institutions.

As a result, the objective of this study is to demonstrate that supervised learning methods can be used to develop credit scoring models from transactional and loan data to determine the creditworthiness of informal merchants. The best performing model is transformed into a business decision-making tool, a credit scorecard, which will help reduce the risk incurred by lenders when giving credit to informal merchants and therefore, increase the access of working capital to support and grow small businesses. The study aims to contribute towards reducing the financial gap in the informal African economy.

To predict the default behaviour of informal merchants in Lesotho, the study used two supervised learning methods: logistic regression and support vector machines. Six (6) logistic regression models and twelve (12) support vector machine models were evaluated for their default predictive power. Support vector machine models outperformed logistic regression models based on the average performance on the Gini coefficient metric. However the case, the best performing model is a logistic regression model with a merchant's credit history as the only feature, which resulted in a 0.6143 Gini coefficient.

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

1 Introduction

1.1 Research Motivation

In collaboration with Nomanini, a Financial Technology company in Cape Town, this research aims to experiment on different supervised learning methods to predict the default risk of small business owners in the informal market using their daily transactional and loan data. The models developed in the study will help financial institutions to make better informed decisions on giving credit to the informal merchants.

1.2 Background of Study

Borrowing and lending have a long history associated with human behavior [26]. Despite the fact that credit has been around since around 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 [15]. Financial institutions create credit scorecards based on information gathered from credit applicants [16] [26].

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, lowering the risk of credit applicants defaulting [1].

1.3 Objectives of Research

Data is an essential component in assisting financial institutions in making credit decisions on credit applicants. As a result, the study will employ anonymized transactional and loan data to carry out the following research objectives:

- Implement a traditional statistical method such as logistic regression and compare the results with machine learning algorithms - support vector machines to predict default of credit applicants.
- Evaluate the performance of the machine learning models.
- Recommend the best model to use to predict the default behaviour of credit applicants.

1.4 Limitations of 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 defaults 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 be able to grow, from being able to carry out everyday transactions more easily to financing expansion. The financial services required by MSMEs differ significantly according to their size, degree of formality, and growth stage – but mostly will require 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 reduces the risk assessment and taking of security among higher-risk MSMEs. Some banks are increasingly making use of their own credit scorecards as an input to the lending decision process, meaning that reliance on traditional collateral is being reduced and working capital financing is increasingly being provided on an unsecured basis [7]. However, this method 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, integrating payments, working capital and data analytics to unlock the latent po-

tential of African's economy. Nomanini operates by connecting any mobile device into a retail point of sale solution for informal merchants and connected to an interoperable merchant wallet.

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. Also, 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.

Furthermore, with data analytics, Nomanini extends working capital loans to merchants through distributors, which allows them to grow their businesses. As a result, an increased volume of goods and services set against reduced operational friction and increased the profits of all participants.

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

Therefore, any solution which requires to solve the challenges of reliance of cash, lack of working capital, and high transactional costs should start with the merchant-first perspective. Due to the complexity and fragmentation of the eco-system means that partnerships between financial technology companies and incumbents such as

banks, mobile and traditional payment companies are key to building such propositions. The Nomanini Fintech platform – a managed cloud solution – connects service providers such as banks and Fast Moving Consumer Goods (FMCG) distributors to the merchants in the informal economy, bringing unity to the landscape 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 [21]. 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 [20]. 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 [19]. 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 [20].

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 between the digital and analogue, financially included and excluded, urban and rural lead to complicated efforts to reach scale and reduce the friction in the supply chain.

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, it 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 either judgemental or statistical [25]. Judgemental scoring depends on an expert to provide a qualitative judgement, whilst a statistical approach depends on quantified characteristics, 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 [25].

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 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 behaviour 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 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].

2.2.2.1 Logistic regression

The first person to use logistic regression in financial risk was Ohlson, where he applied it to predict bankruptcy [22]. 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 a binary logistic regression is either 1 or 0 for, 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 ($i=1, \dots, n$). 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 of as follows:

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

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

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

where the probabilities are restricted to be between 0 and 1, $0 \leq \theta \leq 1$.

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

$$P(D=1|X) = \frac{\sum_{p=0}^k \exp(\beta_p X_p)}{1 + \sum_{p=0}^k \exp(\beta_p X_p)} \quad (3)$$

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:

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

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.

2.2.2.2 Decision trees

Decision trees are recursive partitioning algorithms that develop a tree-like structure representing patterns in an underlying data set [9]. In a given dataset of several credit applicants described by p attributes or characteristics: $x_1, x_2, x_3, \dots, x_p$ which categorised into two groups; defaults and not defaults. The algorithm begins with a root node containing a sample of defaults and not defaults applicants and then, loops over all possible binary splits in order to find the attribute x_p and corresponding cutoff value c_p which gives the best separation into one side having mostly defaults and the other of not defaults [6].

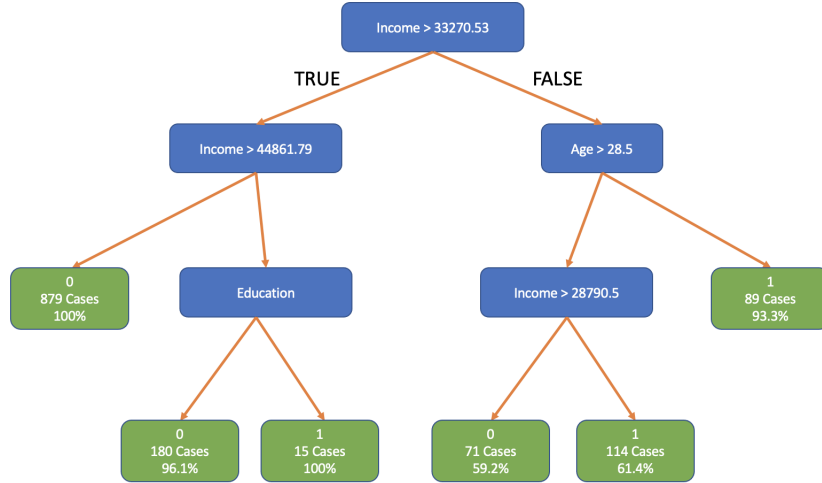


Figure 1: An example of a decision tree

For example, in Figure 1 the figure of merit is optimized when the data in the root

node is split between instances with the income attribute at 33270.53 cut-off point. This procedure is then repeated for the child nodes until a stopping criterion is satisfied. The purity p of a node is the fraction of good instances in the node, and the splitting attribute and cut-off value those that minimize the Gini index $p(1 - p)$ of the created child nodes. The parent node is not split, if the sum of the Gini indices of the child nodes is higher than the Gini index of the parent node, for any attribute or cut-off value [6]. Minimizing the Gini index results into child nodes which are more homogeneous than the parent nodes since, the Gini index is a measure of a statistical dispersion or diversity of population in a node.

In Figure 1, unsplit nodes are depicted by green rectangles and are known as "terminal nodes". Terminal nodes are classified according to the class most prevalent in them. Most often the trees are very large, since a decision tree can be grown until all terminal nodes contain only good credit instances or only bad credit instances, however such tree will be an over-fit of the data. In such situations the performance of the tree can be generalized by technique called pruning, which reduces the size of the decision tree by removing parts of the tree that do not add prediction power [8].

Since the 1980's decision trees have been used to develop credit score models [11], however the limitation of decision trees is its instability from fluctuations in the data sample which may result to large variations in the choice of features at each node and classifications assigned to the instances [6].

2.2.2.3 Support vector machines (SVMs)

SVMs is a popular supervised learning method which can be used without prior domain knowledge of the problem [24]. In the context of statistical learning theory,

Vapnik (1998) was the first to introduce the method [27]. SVMs seek for an optimal hyperplane which constructs a maximum margin separator that creates a decision boundary with the largest distance to example points [12]. This enables a good generalization of the model on the data given.

For example, in Figure 2 (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 2 (b).

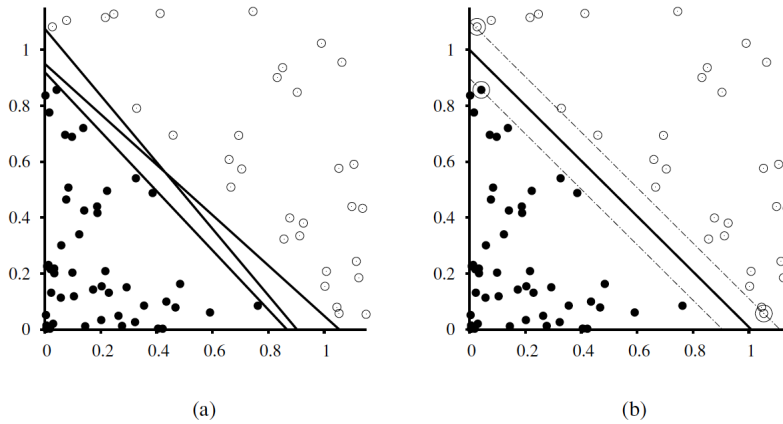


Figure 2: Figure 2(a) shows two classes of points (black and white circles) and three candidate linear separators. Figure 2(b) shows the maximum margin separator (heavy line) is at the midpoint of the margin between dashed lines. The support vectors are the points with large circles on the dashed lines as the closest examples to the separator[24].

In addition, SVMs create a linear separating hyperplane and have the ability to embed data in a higher-dimensional space using the so-called kernel trick for data

that is not linearly separable in the original dimensional space [24]. Hence, data that are not separable in the original space are easily separable in a higher dimension space. For example Figure 3 (a), the data points are not linearly separable in a two-dimensional space but when mapped into a higher dimension space as shown in Figure 3 (b), they become linearly separable.

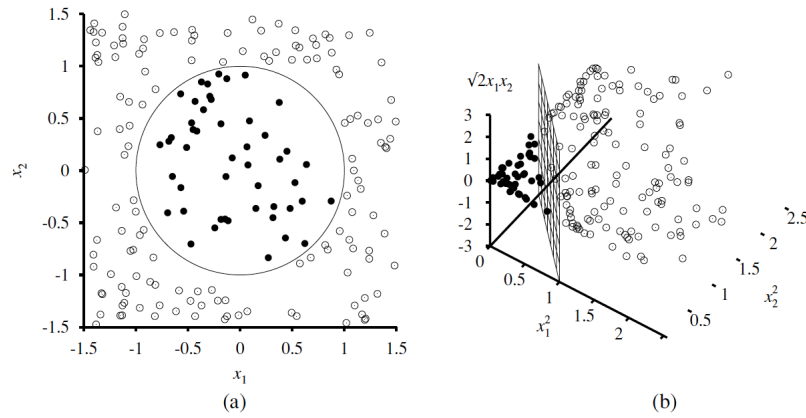


Figure 3: Figure 3(a) shows a two-dimensional training dataset of black and white circles with a decision boundary of $x_1^2 + x_2^2 \leq 1$. Figure 3(b) Shows the same training dataset after mapping it into a three-dimensional space $(x_1^2, x_2^2, \sqrt{2}x_1x_2)$. The circular boundary in Figure 3(a) is transformed into a linear decision boundary in Figure 3(b) [24].

Furthermore, SVMs are a non-parametric method that retain and store training examples and thus, are able to combine the advantages of parametric and non-parametric models [24]. Hence, they can represent complex functions and are more resistant to over-fitting due to better generalization.

2.2.2.4 Other classification methods

Other classification methods developed to build credit scorecards include discriminant analysis, neural networks and ensemble methods such as bagging, boosting, and

random forests [2]. Even though these other techniques are considered to be more powerful and provide better prediction results, they produce very complex models that make them not useful in building credit scoring models where interpretability is a key concern [18].

According to Hand (2006), the potential performance improvements attained by complex models are small compared to the predictive power of simple models, and were often offset by other sources of uncertainty that were exacerbated by the added complexity [14]. Hence in practise the best methodology takes into account costs and benefits, and that is to the reason why logistic regression is the industry standard because the complex models increase implementation costs with a small marginal benefit.

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

2.2.3.1 Confusion matrix

A confusion matrix is a table layout that is used to evaluate the performance of a

classification algorithm. As shown in Table 2 where 0 represents a not default status and 1 represents a default status, the matrix shows two errors of different natures and consequences for the business:

- False Positive - Predicting a bad loan applicant (1) while it is actually good loan applicant (0). The potential loss to the business is interests, principal and to recover the loan costs and amount.
- False Negative - Predicting a good loan applicant (0) while it 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 natures of good predictions that benefit the business as follows:

- True Negative - Predicting a good loan applicant (0) while it 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) while it is in fact a bad loan applicant (1). The potential benefits gained for the business is 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 5, 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.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

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 6 shows the formula that calculates the precision performance metric of a classification algorithm.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

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 7 shows the formula for calculates the recall performance metric of a classification algorithm.

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

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 8 shows the formula for calculates the recall performance metric of a classification algorithm.

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

F-Score metric combines the precision and recall metric performances as shown in Equation 9. This performance metric is used to obtain a balance between precision and recall.

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

2.2.3.2 Receiving operating characteristic (ROC) curve, area under the curve (AUC) and Gini coefficient

The ROC, AUC and Gini coefficient are good metrics to use for classification model evaluations. The ROC Curve is a graphic used to simultaneously show the two types of errors for all possible cut-off probability thresholds of classification [17]. 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 correct identify actual positives (defaults). As shown in Equation 11,

the negative positive rates is the proportion of false positives that were incorrectly classified as defaults to the sum of false positives and true negatives.

$$\text{TruePositiveRate} = \frac{TP}{TP + FN} \quad (10)$$

$$\text{NegativePositiveRate} = \frac{FP}{FP + TN} \quad (11)$$

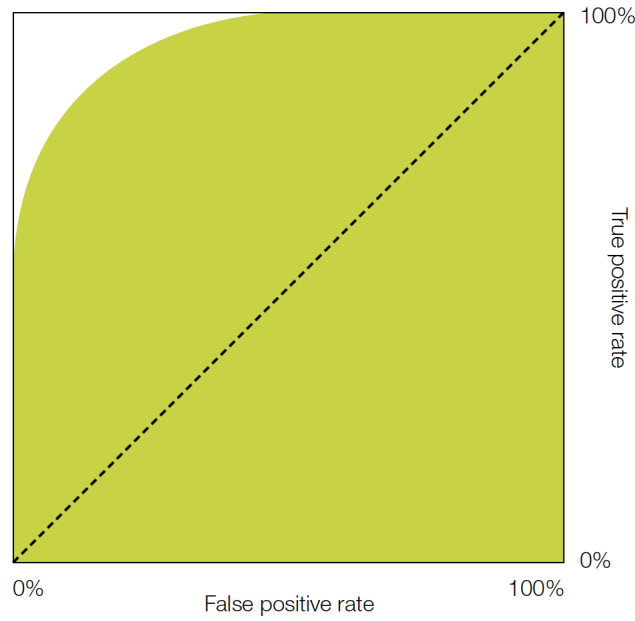


Figure 4: Example of a ROC curve [28].

In Figure 4, the dotted line shows if the model was good at predicting default at random [28]. The AUC measures the overall performance of the classifier of all the thresholds as an area under the ROC curve [17]. The area is measured as a percentage of the box in green in Figure 4. The Gini coefficient is the linear transformation of the AUC, that is a scale of the predictive power of the model.

Please see Equation 12, 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 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].

$$Gini = 2 \times AUC - 1 \tag{12}$$

2.2.3.3 Kolmogorov-Smirnov (KS) Test

The KS test measures the maximum vertical separation between two cumulative curves (defaults and non-defaults) 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 [13].

For example in Figure 5, the maximum difference in the accumulated rates happen at 67% of defaulters and 23% of non-defaulters, making the KS 44%. 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. Therefore, this chapter discussed the financing gap that exists and how innovative technological solutions can be used to solve the

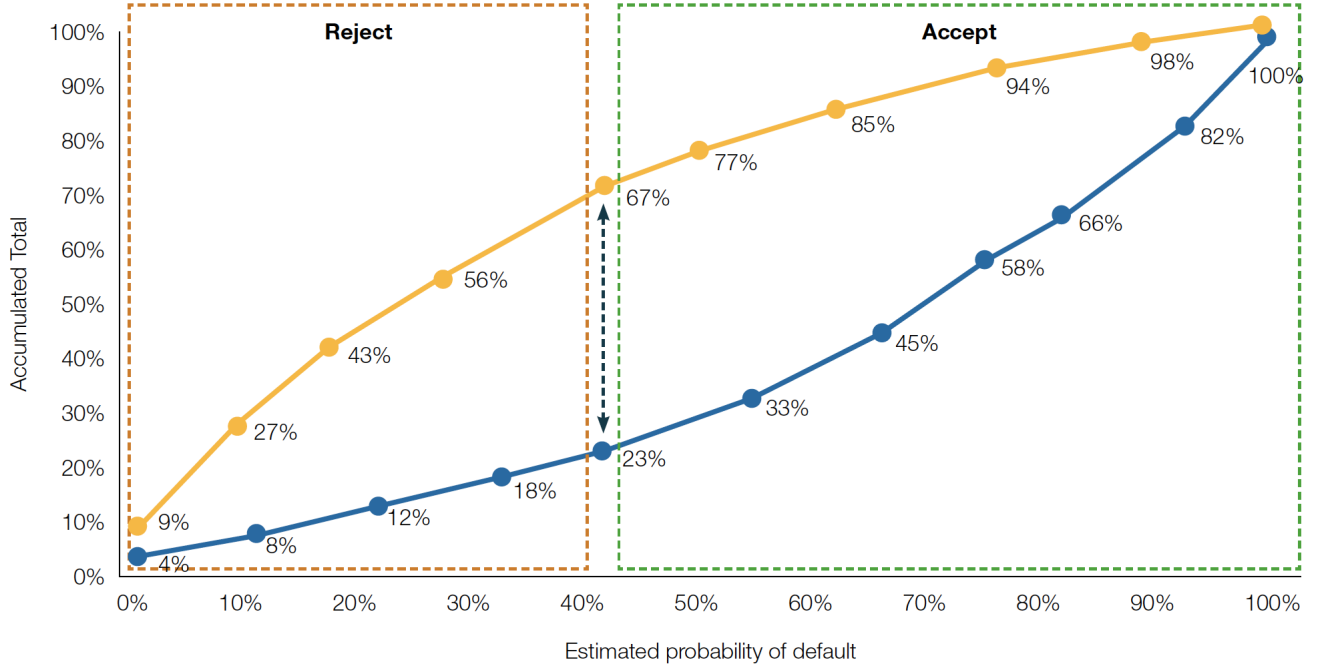


Figure 5: Example of a Kolmogorov-Smimov curve [28].

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 data set consists of anonymized transaction and loan merchant data from No-manini's operations in Lesotho between April and November 2020.

3.1.1 Transaction data

The transaction dataset consist 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 transaction 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.
- No. of airtime pins - The number of times a merchant bought airtime on the platform.
- No. of loan disbursements - The number of times merchant was loaned out money.
- No. of merchant deposits - The number of times a merchant made a deposit to the platform.
- No. 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 of 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 to aggregated information of the number of defaults, delinquents, and paid back loans for each merchant in order to extract valuable insights that will determine the credit behavior of merchants. As a result, the loans dataset was transformed with 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.
- No. of paid loans - The number of times a merchant repaid back a loan on time.
- No. of delinquent loans - The number of times a merchant repaid back a loan within 24hours after collection time.
- No. 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
- No. of airtime pins
- No. of loan disbursements
- No. of merchant deposit
- No. of loan repayments
- Amount
- No. of paid loans
- No. of delinquent loans
- No. of default loans
- Default status

The predictive features were based on transactional and loan data from April 2020 to

October 2020, with the target outcome being merchant default status in November 2020.

3.3 Feature engineering

3.3.1 Average Amount Per Transaction (AAPT) feature

As shown in Equation 13, 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.

$$AAPT = \frac{Amount}{No.ofairtimepins + No.ofloandisbursement + No.ofmerchantdeposits + No.ofloanrepayments} \quad (13)$$

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.

$$credithistory = \frac{No.ofpaidloans}{No.ofpaidloans + No.ofdelinquentloans + No.ofdefaultloans} \quad (14)$$

3.4 Exploratory data analysis

3.4.1 Correlation plot

Figure 6 depicts the correlation of the training set's features. The plot will help to better understand the relationship between the features and will 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.

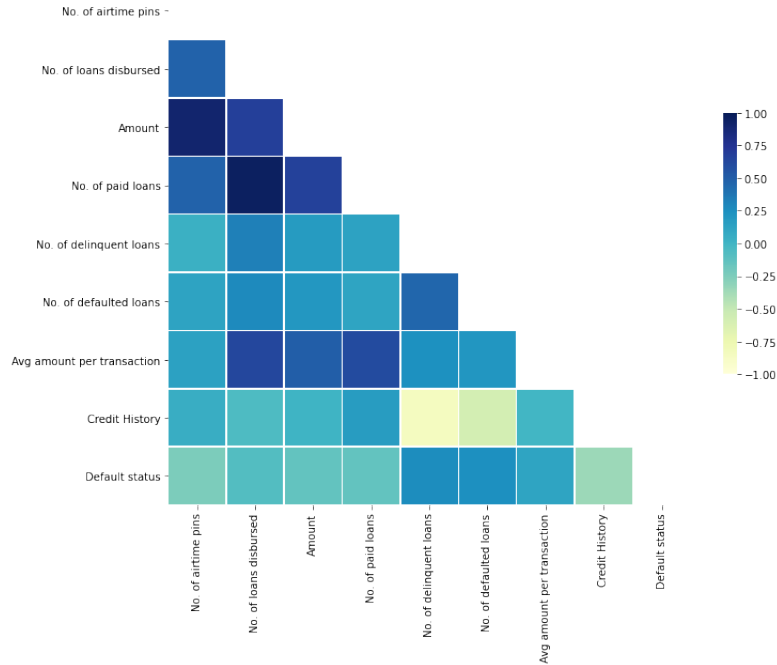


Figure 6: A correlation plot of the features

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 the features is calculated by lowering the 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 [17].

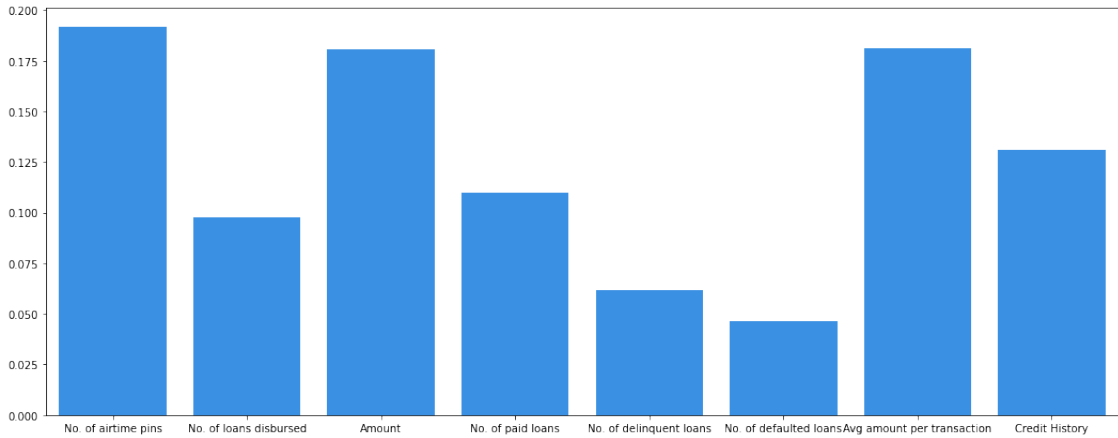


Figure 7: A bar plot showing feature Gini importance

According to Figure 7, the number of airtime pins, amount, and AAPT features were the most important, while the number of defaulted loans and number of delinquent loans features 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

3.4.3.1 Credit history and default status

The credit history feature is a formula that identifies the percentage pay-back rate of a merchants' loans, and the default status feature describes the members who actually defaulted in November. The following findings from the box plot of credit history versus default status in Figure 8:

- The interquartile range (IQR) for defaulted merchants is 47%, while the IQR for non-defaults is 20%. This indicates that the defaulted merchants have a wide distribution spread when compared to the credit history feature, indicating a high variability.
- A lender is more confident to give loans to any merchant with a Credit History of greater than 80%.
- A merchant with less than a 75% credit history is more likely to default on their November loan. This is evident in Figure 8, where more than half of the defaults are located in that range. However the case, there is an outlier

merchant with a 37% credit history who did not default on their November loans.

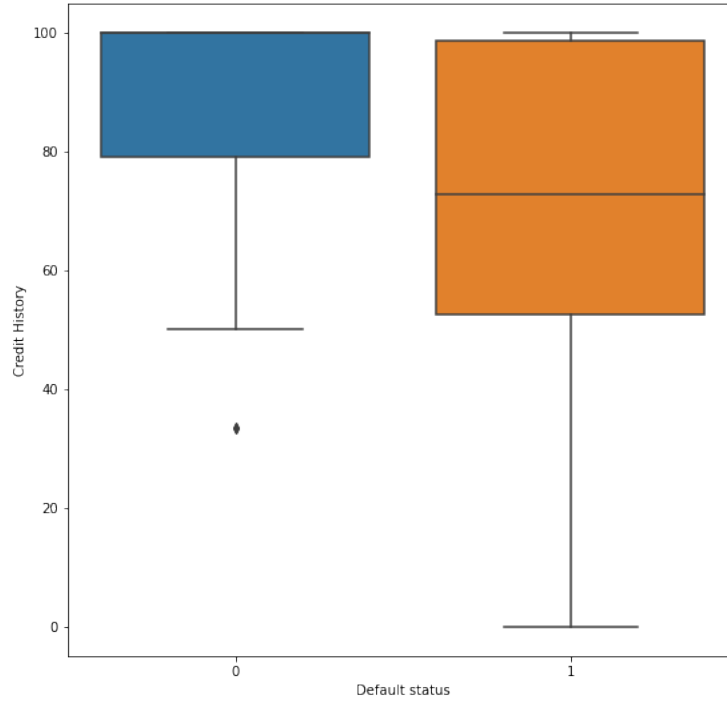


Figure 8: A box plot of Credit History feature against Default Status feature (non-defaulters in blue and defaulters in orange))

3.4.3.2 Average amount per transaction (AAPT) and default status

A box plot of the average amount per transaction (AAPT) feature versus the default status feature demonstrates how the value of a merchant's transaction relates to default behavior. As a result of Figure 9, the following findings can be drawn:

- Non-defaulters have an IQR of 5.5, while defaulters have an IQR of 7, with a median of 16 and 18, respectively. As a result, when a merchant's AAPT is greater than 18, they are more likely to default, even though this is not always the case.

- Because the boxplot shows an overlap in distribution range for defaulters and non-defaulters, the average amount per transaction (AAPT) feature alone is not a good indicator of merchant default status.

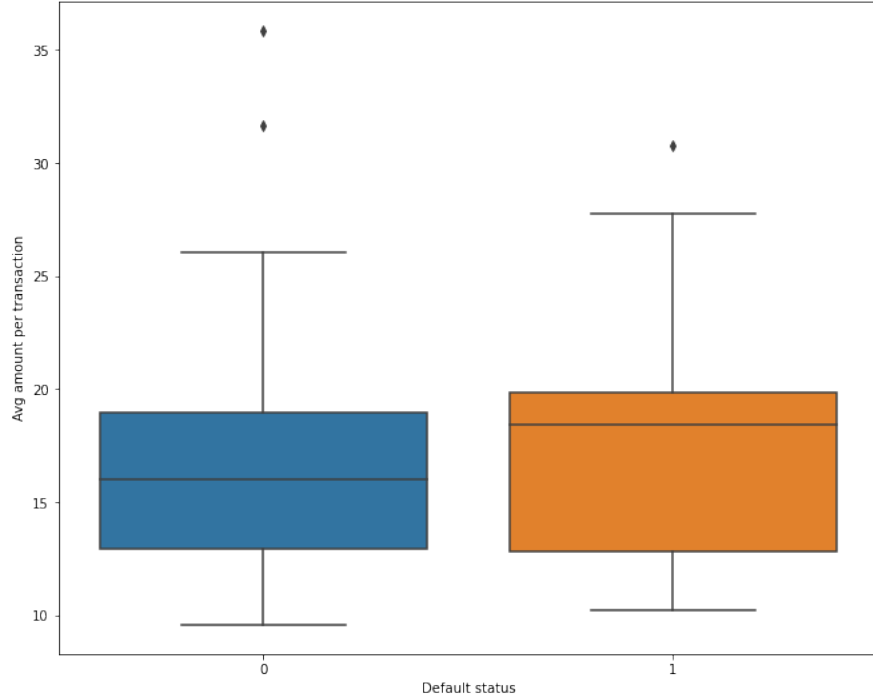


Figure 9: A box plots of Average Amount Per Transaction feature against Default Status (non-defaulters in blue and defaulters in orange)

3.4.3.3 Credit History, Average amount per transaction (AAPT) and default status

Figure 10 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 defaulted merchants 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 6.

- According to Section 3.4.3.1, a merchant with a credit history of less than 80 is more likely to default. However, as shown in Figure 10, 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.

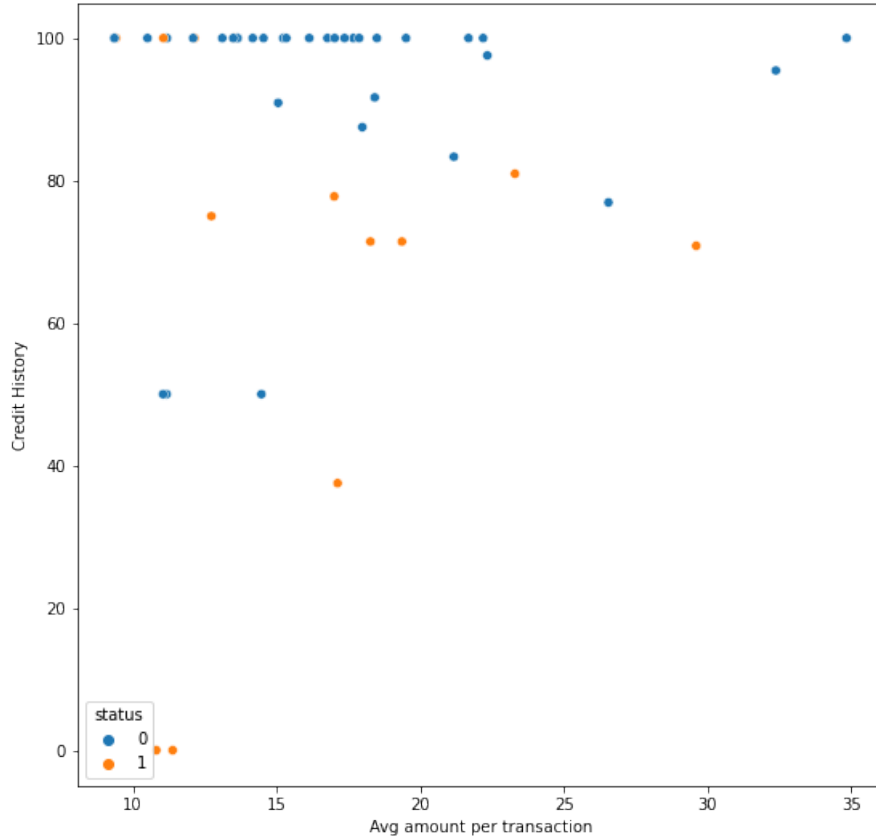


Figure 10: Scatter plot showing the distribution of non-defaulters (blue) and defaulters (orange) when Credit History feature is plotted against Average Amount Per Transaction feature

3.4.4 Class distribution of the target outcome

In the exploratory data analysis section, it is important to investigate for a class imbalance of the target outcome that can impact our model predictions. From Figure 11, the dataset has a 71%-29% split of non-defaults to defaults respectively.

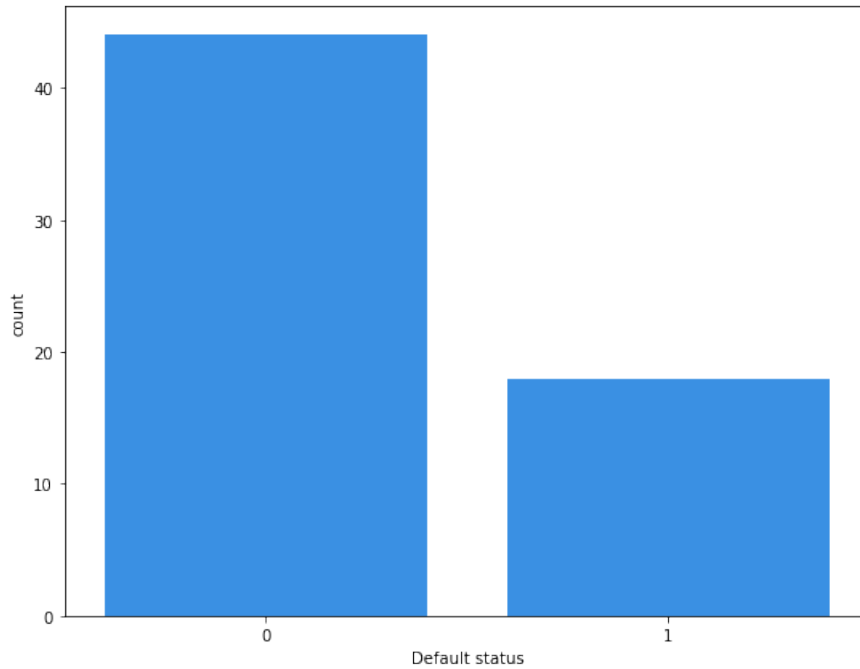


Figure 11: Bar plot showing the class distribution of the target outcome

3.5 Data pre-processing techniques

The following data preprocessing techniques were applied to the dataset:

- No. of merchant deposits and No. 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 split.

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 modeling 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

4.1 Feature selection

There are various statistical procedures used for selecting features that have a statistical significance with the target response [17]. 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 $2^7/128$ 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 No.	Features
Combination 1	No. of delinquent loans
Combination 2	Credit history
Combination 3	Average Amount Per Transaction & Credit history
Combination 4	No. of delinquent loans & Credit history
Combination 5	No. of paid loans & No.of 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 [17]. 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 onto 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.

Model No. 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)
No. of airtime pins						0.7678 (0.6126)
No. of delinquent loans	0.6640 (0.0578)			0.4888 (0.1929)	0.8294 (0.0439)	0.5637 (0.4904)
No. of loan disbursements						-1.5011 (0.7273)
No. 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

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 right 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 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 probabilities 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 [17]. 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 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 cost, resulting in low bias and high variance [23].

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 aren't 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 gamma because it's a non-linear separator. The higher the gamma value is the more the SVM tries to fit the training data. Thus, a high gamma results to high variance and low bias [23].

The grid-search method was used to generate the cost and gamma values in Table 6, with cost ranging from 0.1 to 100 in 0.5 steps and gamma values of 0.1, 1, 10 and 100 were used.

Model No.	Kernel	Cost	Gamma	No. 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 modeling 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, modeling 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 on 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 the Appendix provide a full description of the Table 7 prediction results.

The LR6 model, which combines all features, has the highest test accuracy but a poor cross-validation accuracy, demonstrating its limitations in predicting unseen data. Table 7 shows that the cross-validation accuracy and Gini coefficient metrics of single-featured models, LR1 and LR2, have the best performance. This is confirmed

by the feature selection approach described in Section 4.1 and illustrated in Table 4, where the AIC values for LR1 and LR2 are the lowest.

The LR3 model, on the other hand, which has a low AIC score but is seen to have high training, cross-validation, and test accuracy, has a low Gini coefficient, resulting in poor prediction power on unseen data.

Model No.	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

As a result of the metrics provided in Table 7, the LR2 model is the best performer. This model outperformed the others in cross-validation accuracy and Gini coefficient metrics, demonstrating it is the best model for data generalization.

5.2 Support vector machines

Tables 28 to 52 in the Appendix provide a full description of the Table 8 prediction results.

On the basis of the test accuracy and the Gini coefficient metrics, the support vector machines models in Table 8 reveal that models with linear kernels performed better than models with radial kernels. Models SVM 4B and SVM 1B have the best cross-validation accuracies, but on the other hand, have low test accuracy and an averagely good Gini coefficient, resulting in an inconsistent performance on the evaluation metrics.

With a Gini coefficient of 0.6000, the all-feature combination model SVM 6A with a linear kernel has the greatest Gini coefficient.

Model No.	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 summarized the results of the logistic regression and support vector machines models described in Section 4. The models were evaluated with four evaluation metrics; train accuracy, leave-one-out cross-validation (LOOCV) accuracy, test accuracy and the Gini coefficient.

According to the Gini coefficient metric results, the support vector machines models performed better, with an average Gini coefficient of 0.4499 compared to the logistic regression models' 0.4427 average Gini coefficient, as calculated from Tables 7 and 8. Therefore, from this section the best performing model is the logistic regression model LR2, which has a leave-one-out cross-validation accuracy of 0.7258 and a Gini coefficient of 0.6143.

As a result, the following section will provide an in-depth analysis of the LR2 model's

performance in predicting default as well as the provision of credit-worthiness for informal merchants based on the LR2 model.

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	0%
71% - 80%	2	3	67%
81% - 90%	2	2	100%
91% - 100%	0	0	0%

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

Figure 12 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 12, 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 defaulted merchants exceeds the number of non-defaults, probability of default ranges of 51% - 100% provide a strong evidence for rejecting all loan applications that fall within that range.

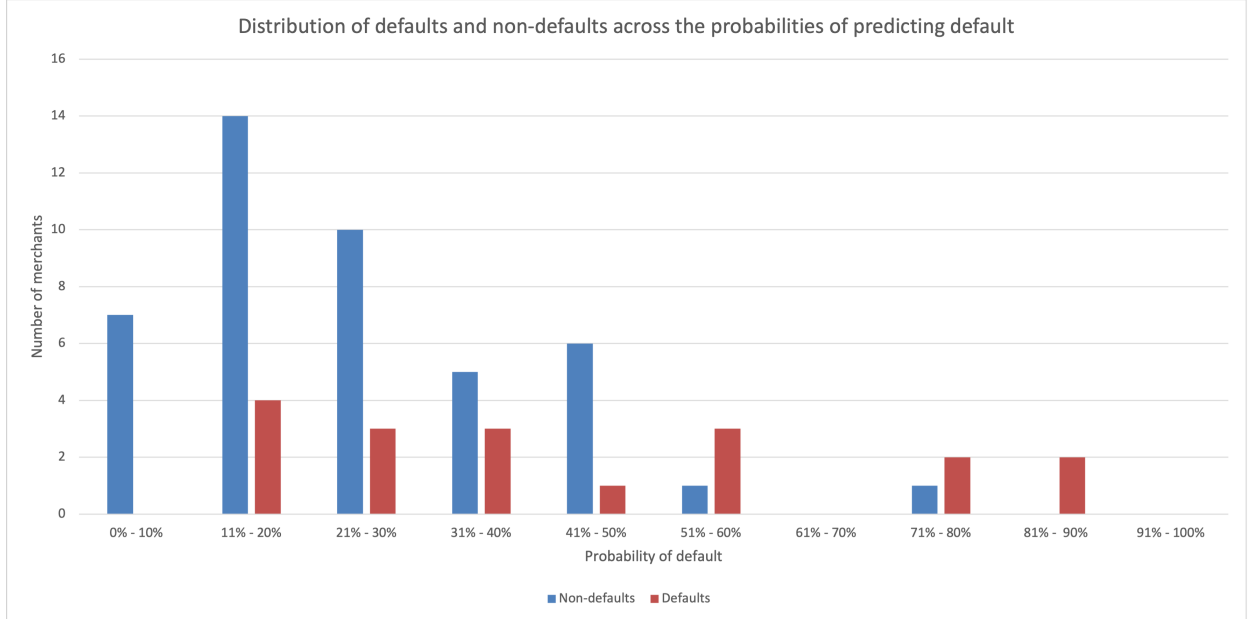


Figure 12: A bar plot showing the distribution of defaulters and non-defaulters as credit scoring model

As a result, once the credit score model has been validated and analyzed, it is

transformed into a business decision-making tool. Based on the data, modeling, and analysis performed on Nomanini's merchants in Lesotho, the results are shown in Figure 13, a credit decision tool that uses a merchant's credit history to quantify the probability of default.

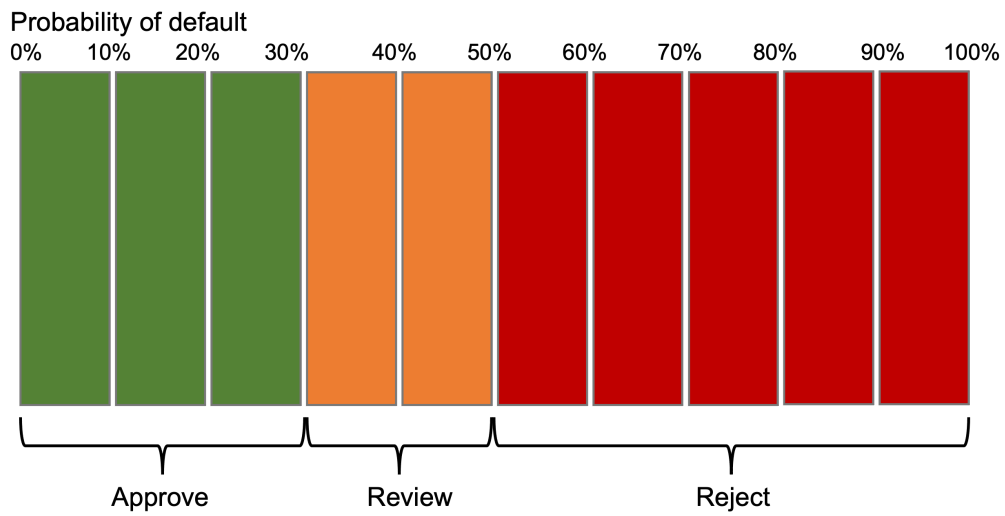


Figure 13: A figure showing the credit decision analysis from the probability of default of the credit scoring model

7 Conclusion

The study developed credit scoring models that predict default by using transactional and loan data from informal merchants. Seven (7) predictive features were extracted and processed from the data. Following that, an Akaike Information Criterion (AIC) feature selection method was used to obtain the five (5) feature combinations that were selected for modeling and analysis. Furthermore, a sixth feature combination of all seven (7) features was included.

The two modeling methods used in the study were logistic regression and support vector machines. Six (6) logistic regression models and twelve (12) support vector machine models were evaluated for their default predictive power on informal merchants. Support vector machine models outperformed logistic regression models based on the average performance on the Gini coefficient metric. The best performing model, however, was a logistic regression model with a merchant's credit history as the only feature, which resulted in a 0.6143 Gini coefficient. The logistic regression method is the most popular among financial institutions due to the ease with which model parameters on how they affect default prediction can be understood and interpreted.

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 modeling 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 modeling 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 modeling 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] Concepción Bartual Sanfeliu, Fernando García García, Francisco Guijarro, and Agustin Romero Civera. Probability of default using the logit model: The impact of explanatory variable and data base selection. In *International Scientific Conference: Whither our Economics*, pages 118–124. Mykolas Romeris University, 2012.
- [6] Joao Bastos. Credit scoring with boosted decision trees. 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] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. Classification and regression trees. 1983.
- [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] Halina Frydman, Edward I Altman, and Duen-Li Kao. Introducing recursive partitioning for financial classification: The case of financial distress. *Journal of Finance*, 40(1):269–91, 1985.
- [12] Rui Ying Goh and Lai Soon Lee. Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research*, 2019, 2019.
- [13] Siana Halim and Yuliana Vina Humira. Credit scoring modeling. *Jurnal Teknik Industri*, 1(N0. 1):17–24, 2014.
- [14] David J Hand. Classifier technology and the illusion of progress. *Statistical science*, pages 1–14, 2006.
- [15] 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.
- [16] David J Hand and Saul D Jacka. *Statistics in finance*. John Wiley & Sons, 1998.
- [17] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

- [18] 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):1466–1476, 2007.
- [19] Leandro Medina, Mr Andrew W Jonelis, and Mehmet Cangul. *The informal economy in Sub-Saharan Africa: Size and determinants*. International Monetary Fund, 2017.
- [20] Nomanini. *FINTECH THE INFORMAL RETAIL ECONOMY IN AFRICA*. Nomanini, 2019.
- [21] Nomanini. *SUPPLY CHAIN FINANCING FOR INFORMAL MSMEs*. 2019.
- [22] James A Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131, 1980.
- [23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [24] Stuart Russell and Peter Norvig. Ai a modern approach. *Learning*, 2(3):4, 2005.
- [25] Mark Schreiner. Scoring: the next breakthrough in microcredit. *Occasional paper*, 7, 2003.
- [26] L. C. Thomas. *Credit scoring and its applications*. SIAM monographs on mathematical modeling and computation. Society for Industrial and Applied Mathematics, Philadelphia, Pa.

- [27] Vladimir Vapnik. The support vector method of function estimation. In *Non-linear modeling*, pages 55–85. Springer, 1998.
- [28] Maria Fernandez Vidal and Fernando Barbon. Credit scoring in financial inclusion. 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] Qingfen Zhang. Modeling the probability of mortgage default via logistic regression and survival analysis. 2015.

8 Appendix

8.1 Logistic regression tables

Feature	Estimate	Standard Error	Z-Value	$\Pr(> z)$
(Intercept)	-0.9067	0.3525	-2.572	0.0101
No. 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
No. 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
No. of paid loans	-0.7121	0.7219	-0.986	0.3239
No. 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
No. of airtime pins	0.7678	1.5162	0.506	0.6126
No. of loan disbursements	-1.5011	4.3044	-0.349	0.7273
Amount	-1.1392	2.0267	-0.562	0.5740
No. of paid loans	0.7072	4.3920	0.161	0.8721
No. 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

8.2 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

8.3 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	397454beef294ce697dfdbbde0649bb	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	fd5bcd080c5441caccef843042dfc6e	20%

Table 52: Showing the probability of default of merchants