

**BINDURA UNIVERSITY OF SCIENCE EDUCATION  
FACULTY OF SCIENCE AND ENGINEERING  
DEPARTMENT OF STATISTICS AND MATHEMATICS**



**MODELLING CREDIT DEFAULT RISK IN MICROFINANCE: MACHINE  
LEARNING APPROACH**

**AUDREY CHINODYA**

**B192356B**

**A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE  
REQUIREMENTS OF THE BACHELOR OF SCIENCE HONOURS DEGREE  
IN STATISTICS AND FINANCIAL MATHEMATICS**

**SUPERVISED BY MS HLUPO**

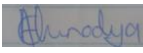
**JUNE 2023**

## APPROVAL FORM

I Chinodya Audrey, do hereby declare that this submission is my work apart from the references of other people's work which has been acknowledged. I do hereby declare that this work has neither been presented in whole or part of any degree at this university.

Student Name

Audrey Chinodya

...  ...

...09/06/2023...

Signature

Date

**Certified by**

Ms Hlupo

...  ...

.....

Supervisor

Signature

Date

Dr.M. MAGODORA

.....

.....

Chairperson

Signature

Date

## **DEDICATION**

I dedicate this dissertation to my family; my father Mr Chinodya, my siblings Mugove, Ashleigh and Almighty Chinodya and my friend Musa Nare. They always believed in the potential I hold and have invested priceless financial, social and time resources to my path to great exploits. Above all glory and praise be to the Almighty God for guidance during my academic career.

## **ACKNOWLEDGEMENTS**

First of almost, I would like to thank the Almighty for leading me through this academic career. Most gratitude goes to my supervisor Ms Hlupo for her untiring supervision. She worked tirelessly as my supervisor, mentor and advisor. I am grateful for her motivation and encouragement.

I would want to express my gratitude to the SFM department staff at large for fostering an environment that allowed me to finish my studies and for providing me with the support that I needed to excel in my studies. I would want to convey my gratitude to my classmates for the time we shared together. Lastly I would like to thank my family for supporting me throughout this academic journey.

## **ABSTRACT**

Credit default risk is the risk of loss that microfinance institutions face when a borrower fails to meet their financial obligations, such as repaying a loan or making interest payments. The goal of this study is to model credit default risk using machine learning models and to determine which model is best for forecasting credit default risk. Stepwise logistic regression, least absolute shrinkage and selection operator (LASSO), neural network, decision trees, and random forest are the models used in this research study. The study also shows the elements that influence credit risk. The data used was obtained from a microfinance in Zimbabwe for the period 2018-2022. There were 12 variables and 47000 observations in the data. The model's efficiency was assessed using the following metrics: accuracy score, recall score, precision score, F1 score, and AUC value. The research findings highlight the elements that contribute to credit default risk in microfinance institutions, as well as the efficiency of machine learning models in forecasting credit default risk. Based on its strong testing performance, F1 and AUC value. Random Forest is the best model for modelling credit default risk in microfinance institutions.

## TABLE OF CONTENTS

APPROVAL FORM .....	I
DEDICATION .....	II
ACKNOWLEDGEMENTS .....	III
ABSTRACT.....	IV
TABLE OF CONTENTS.....	V
LIST OF TABLES .....	VIII
LIST OF FIGURES .....	IX
LIST OF ACRONYMS .....	X
CHAPTER 1: INTRODUCTION .....	1
1.1 Introduction.....	1
1.2 Background of the study .....	1
1.3 The problem statement.....	3
1.4 Research objectives.....	4
1.5 Research questions.....	4
1.6 Assumption of the study .....	4
1.7 Significance of the study.....	4
1.8 Limitations of the study .....	5
1.9 Delimitations of the study.....	6
1.10 Definition of terms .....	6
1.11 Chapter Conclusion.....	6
CHAPTER 2: LITERATURE REVIEW .....	8
2.0 Introduction.....	8
2.1 Theoretical framework.....	8
2.1.0 Non-performing loans .....	8
2.1.1 Default risk.....	8
2.1.2 Credit risk scoring.....	9
2.2 Machine learning .....	9
2.2.1 Logistic regression model.....	10
2.2.1.0 Types of model building techniques .....	11
2.2.1.1 Variable selection techniques for the logistic regression model.....	12
2.2.1.2 Model diagnostic methods for the logistic regression model .....	13
2.2.1.3 Testing for the of the link function hosmer and lemeshow test.....	14

2.2.2 Neural networks (multi-layer perceptron) .....	15
2.2.3 Decision tree model .....	15
2.2.4 Random Forest Model.....	16
2.2.5 Least Absolute Shrinkage and Selection Operator (LASSO) .....	17
2.3 Performance measures .....	18
2.3.0 Confusion matrix .....	18
2.3.1 Accuracy .....	19
2.3.2 Precision, recall and f1 score measure. (f-measure) .....	19
2.4 Empirical literature .....	19
2.5 Research gap .....	20
2.6. Conceptual framework.....	20
2.7 Chapter conclusion.....	21
<b>CHAPTER 3. RESEARCH METHODOLOGY .....</b>	<b>22</b>
3.1 Introduction.....	22
3.2 Research design .....	22
3.3 Description of variables .....	22
3.3 Data collection .....	26
3.4 Data wrangling and pre-processing .....	27
3.5 Proposed Data Analysis .....	27
3.5.1 Modelling of the data set.....	27
3.5.2 Model diagnosis .....	28
3.5.3 Model selection.....	29
3.5.4 Model Evaluation Techniques .....	29
3.6 Validity and reliability in research.....	30
3.7 Chapter conclusion.....	30
<b>CHAPTER 4: DATA PRESENTATION, ANALYSIS AND DISCUSSION .....</b>	<b>31</b>
4.0 Introduction.....	31
4.1 Data exploration and visualizations .....	31
4.1.1 Identifying missing values .....	31
4.1.2 Understanding the distribution of the target variable: .....	32
4.1.3 Identifying correlations between features .....	32
4.3 Models for credit default risk in microfinance .....	33
4.3.1 Stepwise Logistic model .....	33
4.3.2 Least Absolute Shrinkage and Selection Operator (LASSO) .....	35

4.3.3 Neural Network- multi-layer perceptron .....	37
4.3.4 Decision trees .....	38
4.3.5 Random forests .....	40
4.4. Model analysis and discussion.....	42
4.5 Chapter conclusion.....	44
CHAPTER 5: SUMMARY, CONCLUSION AND RECOMMENDATIONS .....	45
5.1 Introduction.....	45
5.2 Summary of the findings.....	45
5.3 Conclusions.....	45
5.4 Recommendations.....	46
5.5 Chapter conclusion.....	46
REFERENCES .....	47
APPENDICES .....	53



## LIST OF TABLES

Table 2. 1: Confusion Matrix .....	18
Table 3.1: Research Variables.....	22
Table 4.1: Summary of the performance evaluation results of the models.....	42

## LIST OF FIGURES

Figure 4.1: Missing values .....	32
Figure 4.2 : Correlation matrix.....	33
Figure 4.3: ROC curve for stepwise logistic regression .....	34
Figure 4.4: LASSO coefficients path .....	36
Figure 4.5: LASSO ROC curve.....	37
Figure 4.6: NN ROC curve .....	38
Figure 4.7: Decision tree .....	39
Figure 4.8: Decision tree ROC curve .....	40
Figure 4.9: RF Variable impotence plot.....	41
Figure 4.10: RF ROC curve .....	42

## **LIST OF ACRONYMS**

LASSO- Least Absolute Shrinkage and Selection Operator

ML- Machine learning

LR- Logistic Regression

MFI- Microfinance Institutions

NN- Neural Network

BIC- Bayesian information criteria

AIC- Akaike information criteria

LRT- likelihood ratio test

ST- score test

MLP- Multiple Layer Perception

FR- Random Forest

TN - True Negative

FN- False Negative

AUC- Area Under the curve

## **CHAPTER 1: INTRODUCTION**

### **1.1 Introduction**

This chapter focuses on utilizing machine learning to model credit default risk in microfinance. According to the RBZ monetary policy report for 2016, issued in January 2017, microfinance institutions (MFIs) have been recognised as an essential component of financial inclusion, with the ability to provide financial services to previously marginalized and unbanked individuals. With such a high level of economic activity in lending by MFIs, there is also a significant amount of credit default risk, as clients fail to repay. As a result, the focus of this study is on obtaining and assessing the usefulness of a machine learning strategy to minimizing the problem of nonperforming loans, which is credit default risk.

### **1.2 Background of the study**

Credit risk is one of the most serious and well-known risks in the financial markets. It is also known as counter party risk, and it is described as the possibility of a borrower failing to pay their loan obligations under particular terms set by the lending institution, Klein (1994). Microfinance institutions are currently facing greater risk and loss than was expected when loans were made, Muriithi (2013). Rising levels of nonperforming loans are a tendency that not only challenges MFIs' viability and sustainability, but also distracts them from achieving their specified goals, Mota et.al (2017). Nonperforming loans appear as liquidity concerns and low profit margins and are an indicator of a bad financial industry and economy. As a result, credit risk modelling is an essential component of financial risk management. This has resulted in the current situation of very restrictive credit and increasing interest by banks in lowering credit risk losses. Furthermore, there has been a recent development in the Zimbabwean economy of crowdfunding and microfinance institutions where credit risk management is critical to ensure profitability in a competitive market. The ability of a company to distinguish between good and poor customers is critical to its long-term success and performance.

Researchers have struggled to understand the credit risk problem since it is based on asymmetric information, resulting in moral hazard and adverse selection concerns. Casu et.al (2020) acknowledge that all contracts and transactions are information-based and take place during financial intermediation. There could be a variety of concerns, such as not all parties being completely informed or some transactions containing additional information that is not available to all participants. This results in an imbalanced flow of information, making it difficult to enter into financial agreements and perhaps leading to inefficient intermediation. The problem arises after the transactions in moral hazard models, such as when lenders are unable to observe the borrower's actions, which affect their probability of default, whereas adverse selection models are characterized by one side not having the information while the transaction is being performed.

### **Machine learning and credit risk modelling**

Machine learning models are commonly utilized in credit rating. MFIs should investigate machine learning techniques to strengthen their overall credit risk management framework because it has the ability to deliver predicted results. A machine learning model is a mathematical predictive model that improves with more data. Machine learning is a quickly growing subject of computational algorithms that seek to accurately imitate human intelligence by learning from their environment. According to Murphy et al. (2015), they are regarded as the workhorse in the new era of so-called big data. Machine learning algorithms use huge sets of data to discover patterns and generate meaningful suggestions. Credit default risk modelling is yet another field where machine learning may be used to provide analytical value because it has access to a large amount of diverse data. Credit risk modelling is the process of assessing data about a person to determine whether or not that person is to payback a loan. Machines can now reach superhuman performance in a variety of disciplines, including engineering, finance, and many others (Brynjolfsson and McAfee, 2017). These machines are increasingly being used for intelligent jobs that were not previously performed by humans. To minimize these risks and increase their capacity to maximize earnings, financial institutions may use credit rating and credit rationing. Credit scoring models have become widely used in the majority of lending institutions globally to avoid the risks associated with moral hazard and adverse selection. Anderson (2007) defines a credit score as the process of translating a set of relevant facts into a numerical

transform that financial institutions may use to make credit choices. These models classify applicants as outstanding or poor based on criteria such as income, age, and marital status. Borrowing score lowers the cost of loan by lowering the likelihood of default through creditworthiness appraisal and, in some situations, fraud detection. It could additionally be able to monitor current loan accounts and prioritize repayment collection. Before extending loans or lines of credit to businesses or individuals, practically all financial institutions now use some type of credit scoring.

Credit scoring models have been built using standard statistical approaches such as Linear Discriminant Analysis (LDA) or linear regression, with the most commonly used models being the logistic, logit, probit, or tobit. Although the majority of these models are parametric, recent research has looked at non-parametric models such as gradient boosting methods, random forests, and machine learning approaches such as artificial neural networks (ANN). Non-parametric models outperform standard models, according to some studies, such as those by Blanco et.al (2013). According to Nyangena (2019) other researches show that traditional models continue to outperform for example, logistic regression outperformed artificial neural networks (ANNs).

As the number of microfinance and peer-to-peer lending organizations grows in Zimbabwe, more research on non-parametric models that investigate the complex interactions between the elements influencing the likelihood of default is necessary. Because there is little literature addressing Zimbabwe's banking credit market, let alone microfinance and peer-to-peer lending, the business stands to profit from the application of non-parametric approaches. The goal to increase the accuracy of the score for credit choices is the major priority of most lending institutions, and even slight changes might result in future earnings for the organization. This suggests that model selection is crucial. This suggests that the model used is important in determining an institution's performance, and improved models that may help redefine how the credit sector has worked in the past should be carefully considered.

### **1.3 The problem statement**

Nonperforming loans have been increasing in the banking industry. This is an indication that something needs to change, most likely as a result of the checking mechanism used by banks before loan disbursements

#### **1.4 Research objectives**

The objectives listed below have been designed to summarize the approach and purpose of the study in order to determine the needs of the problem statement.

1. To identify the most efficient model for modelling credit risk using data from a Zimbabwean setting.
2. To assess the impact of credit default risk on the performance of microfinance institutions.

#### **1.5 Research questions**

1. Which model is the most efficient for modelling credit default risk.
2. How does credit default risk affect the performance of microfinance institutions

#### **1.6 Assumption of the study**

The study is conducted under the assumption that credit risk is the only factor promoting nonperforming loans in Zimbabwean microfinance firms.

#### **1.7 Significance of the study**

The findings of this study aim to add to the literature in this field while also informing the various stakeholders.

##### **1.7.1 To microfinance organizations**

The research assists organizations in lowering the cost of making problematic loans and the opportunity cost of rejecting credit to otherwise profitable consumers.

### **1.7.2 To the financial institutions**

The research is also important for banks because they need a better credit scoring system to evaluate customers and reduce defaults in light of the rise in nonperforming loans, particularly considering the falling interest rate earnings on loans as a result of market share losses to MFIs and internet lenders. This is where the research findings come in handy, because even a modest percentage increase in the models' prediction can result in significant savings for banks, resulting in profits.

### **1.7.3 For future researchers**

The study also serves as a framework for future research in the same field, which enhance on some of the elements that were not covered in this research.

### **1.7.4 To the researcher**

The research is also designed to enrich the researcher academically and practically.

## **1.8 Limitations of the study**

Limitations in a study are restrictions that may reduce the accuracy of the findings.

The following restrictions apply to this study:

### **1.8.1 Inexperience**

The researcher had no hands-on involvement with the project. The supervision reduces the difficulties encountered by the researcher over the course of the investigation.

### **1.8.2 Confidentiality**

The management and personnel were not at ease with releasing some firm information that was relevant to the investigation. I was provided a supporting letter from the university stating that I was conducting research, therefore they disclosed the data.

### **1.8.3 Time constraint.**

Due to time constraints, the researcher was assisted by colleagues because the process of model construction takes much longer than statistical methods for predicting credit risk.



## **1.9 Delimitations of the study**

These are the parameters for the investigation.

1. The study focuses on one of MFIs in Zimbabwe.
2. The data used is from 2018-2022.
3. The study focuses solely on credit default risk as an internal element that the organization can control.

## **1.10 Definition of terms**

**1. Microfinance:** The providing of small loans (microcredit) to the impoverished in order to assist them in engaging in productive activities or growing very small businesses. The phrase may also refer to a wider range of financial services such as credit, savings, and insurance. Roodman (2012)

**2.Non-performing loan (NPL):** A loan that is in default because the borrower has not completed the regular payments for a certain period of time, according to Segal (2020).

**3.Default risk:** The possibility that a borrower would fail to make monthly payments on their loans as specified in their lending arrangements , Jonhson (2012).

**4.Machine learning:** a study of the design and development of algorithms and procedures that allow computers to learn, Curzon (2012)

**5.Credit scoring:** is a statistical analysis used by creditors and financial institutions to determine an individual's creditworthiness, Samreen (2012)

## **1.11 Chapter Conclusion**

This chapter provides an introduction to the research topic. General information and the relationship between credit risk, non-performing loans, and machine learning approaches were analyzed in respect to the problem statement. The objectives are also stated. The chapter concludes by emphasizing the significance of the study and providing definitions for the terms used in this chapter. To establish what other writers have said, the following chapter conducts a detailed analysis of literature on credit default risk, non-performing loans, and machine learning.

## **CHAPTER 2: LITERATURE REVIEW**

### **2.0 Introduction**

This chapter explore the literature on loan defaults and the variables influencing credit defaults focusing mainly on the objectives of the research mentioned in the previous chapter. It covers literature relating to machine learning models, credit risk review and other approaches which are used in assessing the default risk of loans at the MFIs in Zimbabwe. The gaps in the literature are discussed in this chapter, also analysing the relevant theoretical and empirical research on credit risk modelling worldwide.

### **2.1 Theoretical framework**

#### **2.1.0 Non-performing loans**

Non-performing loans are those that have defaulted or are about to fail owing to nonpayment, according to Warue (2013). According to the RBZ (2016) in its 2016 monetary policy Mangudya (2016), non-performing loans (NPLs) pose a threat to economic stability and growth. Furthermore, the RBZ recognizes that high levels of NPLs that exceed the international benchmark of up to 5% can constitute a threat to stability in finance and economic development, and as such, it underlines the importance of addressing the NPL problem as a way to strengthen the Zimbabwean economy.

#### **2.1.1 Default risk**

According to (Crosbie and Bohn, 2019), default risk refers to the possibility that individuals or businesses may be unable to fulfil the required payments on their debt obligation. Almost every sort of credit extension exposes MFIs to default risk. Almost all forms of loan services open lenders and investors to default risk.

### **2.1.2 Credit risk scoring**

Credit risk scoring is the anticipated likelihood that a new client default or not, as determined by classification. According to Hao (2014), risk scoring is used to assess the creditworthiness of incoming applicants. It analyzes the social, demographic, financial, and other data obtained from loan application forms at the time of application to assess the risks connected with credit requests. Application scoring models assist lenders in determining whether new applicants should be granted credit based on customer factors such as income, education, age, and so on.

## **2.2 Machine learning**

Recently, ML methods have been piloted due to advances in computer power, lower costs, and the urgency of large data sets. Machines have been given the power to increase performance without being told exactly how to do the tasks assigned to them. Machines can attain superhuman performance in fields such as engineering, medicine, finance, and many others (Brynjolfsson and McAfee 2017).

According to Elliot et al (2021), a machine learning model is a file that has been taught to recognize specific types of patterns. You train a model over a set of data, giving it a method to use to reason over and learn from the data set. Once the model has been trained, you can use it to reason over new data and generate predictions about specific data sets. There is a substantial body of literature that applies ML approaches to credit risk situation. A number of publications have noted that the methods provide a flexible and powerful framework for estimating default probability and achieving the best prediction performance. Individual classifiers such as logistic regression (LR), semi parametric and non parametric approaches such as Neural networks (NN), Least absolute shrinkage and selection operator (LASSO), Decision tree, and Random forest are the three primary forms of ML. LR is also known as the classic method for modelling credit default risk. The ML techniques mentioned are explained below.

### 2.2.1 Logistic regression model

A logistic regression model is a statistical strategy used in machine learning to comprehend the relationship between the explained variable and one or more explained variables by employing a logistic function in order to assess the probability of default. Because it is simple to understanding, it is considered that the logistic regression model is the most often used statistical technique in most financial institutions to model credit default risk in MFIs. Hilbe (2009) identified that the logistic model can be classed as either binary or ordinal. The dependent variable in an ordinal regression model is made up of two or more categories. A loan in a financial institution can be defaulted (1) or non-defaulted (0). At banks and microfinance institutions, the binary regression model is employed explicitly to estimate default risk. The logistic regression model can potentially employ the maximum likelihood method to estimate model parameters, according to (Muerer and Tolles 2017).

$$\ln \left[ \frac{p(\text{non-default})}{1-p(\text{non-default})} \right] = \beta_0 + \beta_1 + \beta_2 + \dots + \beta_n X_n$$

Where  $n$  is the number of explanatory variables in the model,  $\beta_0$  is the intercept of the model and  $\beta_i$  is the regression coefficient of the  $i$ th explanatory variable.

The Binary Logistic Regression model has the following assumptions:

- Samples must be larger
- Multi-collinearity is not allowed
- There must be a binary response variable.
- A linear relationship is not assumed between the response variables and the independent variables.
- The categorical variable must be exhaustive and mutually exclusive.

To improve the interpretation of the model's results, the odds ratio is calculated using the formula:

$$\frac{p(\text{non-default})}{1-p(\text{non-default})} = \exp[\beta_0 + \beta_1 + \beta_2 + \dots + \beta_n X_n]$$

If the odd ratio is more than one, it suggests that the event is more likely to occur as the predictor increases. If the odds ratio is less than one, it means that the event is less likely to occur as the predictor increases.

### 2.2.1.0 Types of model building techniques

#### 1. Step-wise regression

The most resilient model building method is stated to be step-wise regression, which is based on the concept of adding and eliminating explanatory models from the model. Stepwise regression reduces a list of potential explanatory factors to a manageable set of the most relevant variables. If Y is assumed to be the dependent variable and X1 is assumed to be the independent variable. The model is then fitted as:

$$y_i = \beta_0 + \epsilon_i$$

The following is a fitted linear regression model for all explanatory variables:

$$y_i = \beta_0 + \beta_r X_{ir} + \epsilon_i$$

For  $i = 1, 2, \dots, n$  and  $r = 1, 2, \dots, n$

The F-statistic is calculated for each fitted model and is as follows:

$$F = \frac{MSR(X_r)}{MSE(X_r)}$$

for  $r = 1, 2, \dots, p-1$  which test if the slope is zero or not zero

The process is done numerous times, and comparisons are conducted between the base model, which does not include the explanatory variable, and the preceding model, which has fewer parameters. Explanatory variables introduced to the model are then removed using this method because they may be irrelevant in following models.

#### 2. Backward selection

This is a standard method for selecting variables in linear regression, and it begins with a base model that contains all possible explanatory variables. The best model is then chosen using the F statistic computed as follows:

$$\frac{MSR(X_1, \dots, X_{p-1})}{MSE(X_1, \dots, X_{p-1})}$$

Where p is the number of explanatory variables in the model.

During the calculation of F, a variable is removed and the F values of the model are compared. A lower value of the F corresponding to the model is chosen, and the variables are removed until the optimal model is identified and chosen. The variables are no longer dropped after identifying and selecting the best model because the best model has already been identified.

### **3. Forward selection method**

The technique also begins with a base model that has no explanatory variables. Explanatory variables are then added to the model, and the model is compared to see if there is a significant difference. In model comparison and assessing the new explanatory variables, the BIC, AIC, and likelihood ratio tests are used.

#### **2.2.1.1 Variable selection techniques for the logistic regression model**

##### **1. The Akaike Information Criteria**

This evaluates the model's badness after adding or eliminating various explanatory variables, as described by Akaike in 1971.

$$AIC = -2\log L(\hat{\beta}) - 2p$$

If the AIC value is lower, the model is said to be better; thus, the lower the number, the better the model.

##### **2. The Bayesian Information Criterion**

BIC is calculated as follows:

$$BIC = -2\log L(\hat{\beta}) + p\log(n)$$

If the BIC value is lower, the model is considered to be better; so the lower the number, the better the model.

##### **3. Likelihood ratio test**

The likelihood ratio test (LRT) is used in the assessment of the variable whenever it has been included into the model, according to Shipe et.al (2019). LRT is a hypothesis that assists in selecting the optimal model from two nested models. The LRT is marked as follows:

$$LR = -2L \log \frac{\text{likelihood without variable}}{\text{likelihood with variable}}$$

#### 4. Score test

The score test is also known as the Cochran-Armitage trend, the Rao test, and the Multiplier test Shipe et.al (2019). This method determines the explanatory variable from the model as well as any changes in the model's significance. The logistic regression model is identified by the formula:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i n_i}{n}$$

The test statistic for the score test (ST) is given as

$$ST = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^n (x_i - \bar{x})^2}}$$

For a significant coefficient  $ST \sim N(0, 1)$

#### 2.2.1.2 Model diagnostic methods for the logistic regression model

- **Pearson chi-square statistic**

Is a method used in the goodness of fit test for the logistic regression model. It is calculated as:

$$X^2 = \sum_{i=1}^n \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)} \sim X^2_{n-p}$$

- **Cox and snell  $r^2$  statistics**

This is a mathematical calculation which is based on the formula given below

$$R^2 = 1 - \left[ \frac{-2 \log L}{-2 \log L_k} \right] 2/n$$

Where  $H_0$  represents the fitted model with the explanatory variable and the kth model has K explanatory variables.  $R^2$  values greater than one are considered better models.

- **Wald test**



This test determines the significance of explanatory factors included in the regression model. It is used to determine whether the explanatory variables in a model are significant. The dependent variable is the one that is changed, as demonstrated by the model's significance.

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Test Statistic is given as:

$$W = \frac{\hat{B}}{SE_B}$$

For the statistic to be considered significant, it must follow the usual normal distribution.

### **2.2.1.3 Testing for the appropriateness of the link function hosmer and lemeshow test**

The Hosmer and Lemeshow test is used to evaluate the model to see if it best matches the data. When a parameter is introduced to the model, the assessment includes looking at the modifications to see whether there is a change from the original model.

The hypothesis tested are:

$$H_0 : E[Y] = \frac{\epsilon^{ni}}{1 + \epsilon^{ni}}$$

$$H_1 : E[Y] \neq \frac{\epsilon^{ni}}{1 + \epsilon^{ni}}$$

Test statistic is given by:

$$C = \sum_{k=1}^g \frac{(y_k - n_k \bar{p}_k)^2}{n_k \bar{p}_k (1 - \bar{p}_k)}$$

This test statistic has g-2 degrees of freedom and follows an X<sup>2</sup> distribution, where g is the number of groups.

### 2.2.2 Neural networks (multi-layer perceptron)

Neural networks (NN) are models based on the flexibility of the human brain to describe any non-linear relationship between explanatory variables and response variables Bishop et.al (1995). The NN model has several architectures but is utilized on multiple layer perception (MLP). All of the neurons from the MLP's explanatory variables are made up of hidden neurons of any number and an output layer, which in our example becomes one neuron. The activation function is used to compute the weighted inputs and the bias term  $b_i^{(1)}$ , resulting in

$$h_i = f(1)b_i + \sum_{j=1} X_{Wij}X_j, \text{ where } j=1$$

Where X is the weighted matrix and  $W_{ij}$  is the link between the input j and the hidden neuron i. In the output layer, the activation function  $f(1)$  is a sigmoid function represented by  $f(2)x$ ; hence,

$$f(2)x = \frac{1}{1+exp^{-x}}$$

As a result, the answer probability is:

$$\pi = f(2)(b_2 + \sum_{j=1} X_{vjh}j), \text{ where } j=1$$

### 2.2.3 Decision tree model

The decision tree model is a classification model that splits the data's feature space into subsets with comparable qualities, therefore classifying or grouping the data sample into homogeneous classes or groups. They are called decision trees because the split of data may be visualized in a tree arrangement. The tree have a root node, a child node, and leaves. The decision tree's root node, which contains the whole population of the research, is at the top. The child nodes are situated at the bottom of the root node and contain a more homogeneous sample produced from the parent node. The leaves, which are used for categorisation, are located at the bottom of the decision trees. According to Matre (2019), one of the most prominent ways for splitting the dataset from the complete population is the entropy measure strategy, which generates the child nodes depending on the information acquired during the split. For a node to evolve, the

information collected from the new node must be greater than the information obtained from the previous node. The entropy of a node is calculated as follows:

$$entropy = -p_1 \log(p_1) - p_2 \log(p_2)$$

where  $p_1$  and  $p_2$  are parent nodes 1 and 2, respectively. The data obtained through the establishment of a child node is then provided by:

$$IG(\text{parent}, \text{children}) = \text{entropy}(\text{parent}) - [p(c_1) \text{entropy}(c_1) + p(c_2) \text{entropy}(c_2) + \dots]$$

where  $c_k$  is the  $k^{\text{th}}$  child node.

#### 2.2.4 Random Forest Model

Many studies have been conducted in order to improve decision trees in general. The random forest was the most significant improvement. It is the work of Geman et.al (1992) who attempted to solve the instability of decision trees. Breiman (2001) finally defined the random forest. Random forests improve the efficiency of decision trees, hence improving predictive power accuracy. Breiman (1996) used the random forest to resolve and address the unstable nature of decision trees like the Classification and Regression trees. It is crucial to first understand the CART in order to properly appreciate the random forest. Breiman introduced CART in 1984. It is a statistical procedure applied on a discrete or continuous dependent variable or output.

A regression tree is a classification tree of the independent variable that is generated by partitioning the multiple dimensional space of independent variables into distinct areas with the assumed constant using the method of recurring relations. A random forest model is more effective than other models because, unlike standard models, it allows fitting models with the dependant variable forced totally on the output. The CART models must be used to solve the problem with high-dimensional features. A recursive method is utilized to estimate the CART parameters using training data by selecting features from the  $x$  ( $x_1, \dots, x_D$ ) as well as the parameters  $L_j$  that reduce the residual sum of squared errors. To avoid overfitting of the training data, the pruning strategy is employed to halt the tree from expanding. Breiman (1996). The RF may be developed using the same principles as the CART and thus have the ability to model a big sample as well as the largest predictions even when  $J \ll N$  while preserving effectiveness and efficiency. The most popular RF class is determined by a community vote process. It

has been observed that the RF can generate predictions using all explanatory variables. The hierarchy of the predictors of the RF is improved, and the RF is ranked using two primary measures. The initial instances of the RF are the mean decrease purity and the mean decrease accuracy. According to Breiman (2001), the mean reduction impurity is given by the following equation.

$$MDA(X_r) = \frac{1}{B} \sum_1^B (e_{OBB} - e_{OBB_{ir}})$$

When analyzing the RF algorithm, Breiman (2001) explains how these mistakes are exploited. One of the RF model distinctive characteristics is its capacity to learn the class imbalance, which allows for improved performance when dealing with credit risk data that contains a lot of imbalances between the input and the output, therefore the response variable.

### **2.2.5 Least Absolute Shrinkage and Selection Operator (LASSO)**

It is a regression analysis machine learning algorithm. It is a sort of linear regression model with regularisation to avoid overfitting. The basic idea underlying LASSO is to add a penalty term to a standard linear regression cost function, which limits the magnitude of the model's coefficients. This penalty term is called the sum of the coefficients' absolute values multiplied by a hyperparameter alpha. We may regulate the amount of shrinkage applied to the coefficients by varying the value of alpha.

With LASSO, you may specify all of the subsets and greedy variations for feature selection and then compare their computing costs. Describe what happens to calculated LASSO coefficients as tuning parameter lambda is modified. The tuning parameters are chosen based on the AIC, BIC, and cross validation prediction errors. Xiang et.al( 2017).

Lasso also has the ability to do feature selection, which means it can automatically choose which features are most useful for predicting the target variable. This is accomplished by effectively deleting additional characteristics from the model by

driving their coefficients to zero. Overall, LASSO is an effective method for developing simple linear models that generalize well to new data.

### 2.3 Performance measures

Data is typically divided into two sets: training and test data, which allows these models to be performed. The training data parameters are then utilized to see how these parameters perform on the test dataset. Although this is not an essential requirement, it is vital that the data be split into two sets or batches with 70/30% splits AI-Shayea et.al (2011). When the model is entirely created, its efficiency on a certain dataset is assessed. The Confusion matrix, Precision, Accuracy, Recall, F1 score, and AUC technique are all used to assess the quality and efficiency of the models.

#### 2.3.0 Confusion matrix

The classification model performance is plotted on a table to create the confusion matrix, which is a table holding the classification model data descriptions. The confusion matrix table can display several parameters. True Negative (TN) refers to the model's accurately predicted negative values based on the actual data. False Negative (FN) numbers are those that were incorrectly forecasted based on the Table.

**Table 2. 1: Confusion Matrix**

	<b>Predicted Condition Positive (1)</b>	<b>Predicted Condition Negative (0)</b>
<b>True Condition = Positive (1)</b>	True Positive (TP) (Correctly predict a true condition)	False Positive (FP) (wrongly predict a wrong negative)
<b>True Condition - Negative (0)</b>	False Negative (FN) (wrong predict a true condition)	True Negative (TN) (correctly predict a negative condition)

Source: **Mercadier, et al. (2019)**.

False Positive (FP) these are the wrongly predicted positive values forecasted by the model using the actual data.

### **2.3.1 Accuracy**

The accuracy of the model is typically used to assess performance. Accuracy is regarded as a weak metric of model evaluation since it does not account for class distribution AI-Shayea et.al (2010).

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

### **2.3.2 Precision, recall and f1 score measure. (f-measure)**

One of the primary issues is the imbalance in credit risk statistics. If the class of interest is significantly outnumbered, the dataset is said to be imbalanced. Liu et al. (2018) provided the following equations for the above measures:

$$\text{Precision} = \frac{TP}{TP+FN}$$

$$\text{F measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

## **2.4 Empirical literature**

According to Wiginton (1980), Durand (1941) conducted one of the earliest studies on credit risk modelling or scoring that used quantitative methodologies on consumer credit risk assessment. Finlay (2011) rated the applications as positive or negative using a quadratic discriminant modelling method. Following that, Ohlson (1980) published a paper in which he established the notion of credit risk analysis through the use of logistic regression modelling. He borrowed White papers, (Jarrow and Turnbull, 1995) and (Santomero and Vinso, 1977) paper. They developed a strategy for predicting failure based on likelihood, which influenced his research. Because of the nature of creditworthiness assessment, adequate research publications demonstrating the performance of commercial consumer credit risk modelling are hard to come by.

The current body of research focuses on two key areas: predicting the downfall of a business and modelling individual credit risk. (Altman and Bland ,1994) used both linear discriminant analysis and neural networks to evaluate over a thousand Italian enterprises for corporate distress. This was one of the papers that looked into corporate risk modelling. Altman (1968) used the first multivariate analysis for the corporate section, known as the Z score model, which is still in use today. He noticed the shortcomings of absolute financial ratio comparisons at the time and offered an extension that combines several measures into meaningful predictive models utilizing statistical techniques given by Fisher (1936) called multiple discriminant analysis (MDA), Altman (1968). Logit and probit analysis first appeared in the late 1970s with the work of Ohlson (1980), who used a less restrictive logistic regression (LR) model, and Zmijewski (1984), who used a similar method to estimate the likelihood of default. Because computational power has increased, machine learning methods have been tested.

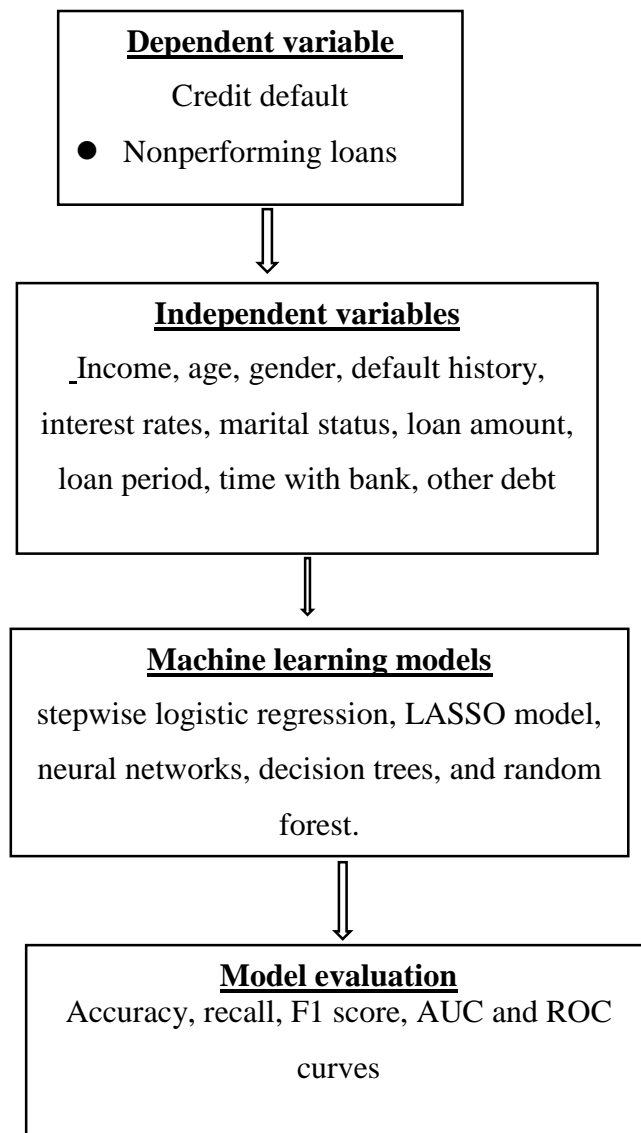
## **2.5 Research gap**

It is disappointing that little to almost no studies have been conducted in Zimbabwe, especially given the country's unstable economy and deteriorating microfinance institutions. As a result, the fact that it is being conducted in Zimbabwe makes this study extremely important. This provides a clear picture of where MFIs stand in the struggle to minimize nonperforming loans. It helps policymakers and other stakeholders in making decisions about how to improve. As a result of technological advancements, this research focuses on how we may better model and anticipate credit default risk using a machine learning technique. This research focuses mainly on logistic regression, LASSO, random forest, decision trees, and artificial neuron network.

## **2.6 Conceptual framework**

Using the conceptual framework, the relationship between credit default and the response variables is demonstrated. It is crucial to find and choose relevant characteristics that are likely to have an impact on credit default in order to construct a successful credit default prediction model. The conceptual framework for modelling

credit default using a machine learning approach involves collecting and analyzing historical data on credit default, choosing relevant characteristics, utilizing machine learning algorithms to develop models, and assessing model performance.



## 2.7 Chapter conclusion

The above chapter expanded on the literature relating to credit risk and the models which can be fitted to predict credit risks at the micro finances in Zimbabwe. The following chapter expand on the methodology used in this research.



## CHAPTER 3. RESEARCH METHODOLOGY

### 3.1 Introduction

This chapter looks into models that are used to analyse the collected data. The areas elaborated are research design, procedures of collecting data, analytical models included and proposed data analysis and presentation. Furthermore, justifications for the variables used in this study are provided, and a chapter summary is provided at the end.

### 3.2 Research design

The research design is the method used to answer the research questions in chapter one. The research design is a plan used to clearly integrate different parts of a study in order to ensure that the research topic is successfully addressed (Burch & Carolyn, 2016). This study used a descriptive research strategy, which helped in gathering relevant information on the specific subject under investigation and describing characteristics of the variables in this research, resulting in numerical data or data that could be translated into meaningful statistics. Descriptive quantitative research employs quantitative data to establish facts and detect trends in research. This study methodology produced unbiased statistical and logical results.

### 3.3 Description of variables

Table 3.1 shows the description of the data which was used in this research study and justification of the selected data variables.

**Table 3.1: Research Variables**

variable	Description of the variable	Justification of the variables
----------	-----------------------------	--------------------------------

term	Duration of the loan for example 12, 24 or 36 months	There is an adverse relationship between payback and repayment period.(Wongnaa and Awunyo-Vitor, 2013) indicated that the longer the repayment period, the lower the default rate and that high repayment leads institutions to cut their interest rate and cost of loan processing.
rate	The interest rate charged	When the interest rate on a loan is high, the cost of borrowing rises. As result, the borrower's capacity to repay the loan reduces since a higher interest rate increases the amount of instalment paid after a specific interval set in the loan's terms and conditions. According to Mansoori (2009), the most important element influencing loan repayment is the interest rate.The interest rate is identified as one of the factors of loan default likelihood (Salas and Saurina, 2002). According to Coravos (2010), high interest rates increase the likelihood of loan default.
Income	Indicates the income of the borrower	The repayment of a debt is closely related to income. Therefore, a borrower's monthly income is believed to play significant roles in loan repayment. When a borrower has a large monthly income from multiple sources, including the utilization of a current loan, the repayment capacity of the borrower is likely to be high. A low monthly income of the borrower, on the

		<p>other hand, result in bad loan performance. According to Abafita (2003), income is a crucial element that improves loan repayment performance. Another study found that wealth contributes positive to a borrower's trustworthiness (Arene, 1993). As a result, the degree of income predicts the likelihood of loan default.</p>
gender	Male or female	<p>It is expected that the correlation between gender(males) and loan default is positive. This is due to the fact that males usually have more duties as family leaders and as a result they may use a loan acquired for investment for other objectives such as fee payment and other utility expenditures. This makes loan repayment harder because no money is generated, increasing the chances of not repaying the loan on time Ibekwe et.al (2007). According to Wongnaa and Awunyo-Vitor (2013), females have a better loan repayment history than males.</p>
Age	Represent the age of the borrower	<p>The performance of a loan is influenced by age. Gan (2012) identifies age as a probable cause of default. Age can have an impact on loan default in both positive and negative ways. According to Awunyo-Vitor (2012), age has a positive effect on loan payback. According to an empirical study conducted by Arene (1993), age has a</p>

		beneficial impact on a borrower's creditworthiness. Individuals who are able to work in the labor markets are provided loans. In Zimbabwe, it runs from 18 to 65 years.
Marital status		According to Zohair (2013), there is a negative correlation between marital status and loan default. That is married couples are more likely to receive help from their partners and as a result loans may be repaid on time. Therefore, the probability of married respondents to fail on loan repayment is lower than peers who are single, separated or widowed and may not have any support from anyone. Duy, V.Q. (2013).
time with bank		According to Haron (2013), the length of time a customer has been with a bank can be a factor in estimating the chance of a loan default. If a customer has had an account with a bank for a long time and has proven good financial behaviour over time, the bank may be more willing to give favourable loan conditions and interest rates.
Loan amount	The loan amount given to the borrower.	According to Mokhtar et.al (2012), loan amount can be a determinant of loan default because it is one factor that affects the borrower's ability to repay the loan. Generally, larger loans require higher monthly payments and may have longer repayment terms. If the borrower has difficulty making these larger

		payments or experiences financial hardship during the repayment period, they may be more likely to default on the loan. Additionally, the loan amount can also be an indicator of the borrower's creditworthiness and risk profile, Orlova (2021)
default history		According to Baklout (2013) Loan default history is often considered a strong determinant of future loan default because it indicates a borrower's past behavior in repaying their debts. If a borrower has a history of defaulting on loans or making late payments, lenders are more likely to see them as a higher risk and may be less willing to extend credit or offer favorable terms. This is because past payment behavior is seen as a predictor of future payment behavior, and lenders want to minimize their risk of losses due to non-payment, carling et.al (2007).

### 3.3 Data collection

Secondary data was gathered for this study. The researcher gathered information from a Microfinance institution in Zimbabwe. The researcher collected data from 2018 to 2022. The data was also obtained under tight restrictions because it was considered confidential. Data collected includes demographic information, credit history information, and other pertinent indicators that may influence the chance of default. For the five years under study, the data collected included 47000 observations and twelve variables. Excel and R were used to process the data and data analysis.

### **3.4 Data wrangling and pre-processing**

Data wrangling and pre-processing are important steps in the data analysis pipeline that involve transforming and preparing raw data for analysis. In this research the following data wrangling and pre-processing tasks were done:

Data cleaning: This involves identifying and fixing errors, missing values, and inconsistencies in the data. Missing values were checked and all the variables has values, duplicate records were also assessed.

Data transformation: This involves translating data into an analytical format. The whole dataset was converted to numerical format.

Overall, data wrangling and pre-processing are essential steps in the data analysis process that can have a significant impact on the accuracy and reliability of the results. By performing these tasks carefully and systematically, the researcher have ensured that the research working with clean and well-formatted data that is suitable for analysis.

### **3.5 Proposed Data Analysis**

#### **3.5.1 Modelling of the data set**

The researcher built a training and test set before beginning the modelling process. The training set was used for modelling, while the test set was utilized to validate the various models employed in the modelling process and to check the accuracy of the data. The following models were then fitted by the researcher:

##### **1. Step wise Model of logistic regression**

It is a statistical method for determining the most essential variables in a logistic regression model. It entails adding and eliminating predictor variables from the model

iteratively based on their relevance until an optimal collection of variables is determined.

## **2. Neural network**

They are a form of machine learning algorithm based on the structure and function of the human brain. They are made up of layers of artificial neurons, which are mathematical functions that take input and output based on a set of learnt parameters.

## **3. Random forest**

It is an ensemble learning technique that combines the predictions of multiple decision trees in order to improve the precision and durability of machine learning models. A random forest model is constructed using a random portion of the training data and a random subset of the input features. By constructing several trees in this manner and integrating their predictions, the random forest model can reduce overfitting and improve generalization performance.

## **4. Decision tree model**

It is a classification and regression analysis machine learning algorithm. It builds a tree-structured model in which each internal node represents a parameter test, each branch represents a test result, and each leaf node represents a class label or numerical value. The decision tree algorithm recursively splits the data based on the values of the input features.

## **5. Least absolute shrinkage and selector operator**

It is a sort of linear regression that use L1 regularization to successfully execute feature selection by shrinking some of the coefficients to zero. This is useful for estimating credit default risk since it allows you to identify which characteristics are most relevant in determining whether a borrower is likely to default on a loan.

### **3.5.2 Model diagnosis**

Model diagnostics are based on the connection function, precision, and overall model sufficiency. Regression diagnostics is a subset of regression analysis whose target is to

determine whether the calculated model, the assumptions made about the data and the model are matched with the recorded data. Its purpose is to assess the model assumptions and determine whether any findings have a large, unfavourable impact on the study. The analysis of variance (ANOVA) is used to confirm if the logistic regression assumptions have been met.

### **3.5.3 Model selection**

This is the process of analyzing the relative importance of many statistical models and finding which one best fits the observed data. There are other approaches for selecting models, however in this study, the researcher employed the Information Criteria Technique. The best model is chosen from the list of models based on information criteria such as Bayesian and Akaike information criteria. The model with the lowest Akaike and Bayesian scores is used.

### **3.5.4 Model Evaluation Techniques**

The Wald test is used to determine the best model's appropriateness. The null hypothesis of the associated coefficient being zero is rejected. If we fail to reject the null hypothesis, we might conclude that the estimators are useful. This technique is used to extract important variables from a group of predictors used in a variety of models using binary or continuous variables.

The Hosmer-Lemeshow test is used to determine the model's accuracy. If the p value is smaller than the set significance level, the model is less significant. It is just a chi-square goodness of fit test for grouped data. The Cox and Snell test, the F distribution, and analysis of variance (ANOVA) is used to assess the overall model's appropriateness.

To assess the prediction capacities of the models, the researcher created a confusion matrix for all of the models used in this study. The confusion matrix displayed real positives, false negatives, false positives, and true negatives. After creating the confusion matrix, the researcher calculated the precision, sensitivity, accuracy,



specificity, error rate, ROC AUC, and F1-score for comparison reasons. To eliminate multicollinearity in the logistic regression model, the researcher generated variance inflation factors for the variables in the fitted model.

### **3.6 Validity and reliability in research**

The principles of reliability and validity are used to measure research quality. A measure's consistency is referred to as its reliability, whereas its degree of correctness is referred to as its validity. The utilization of secondary data from a well-known microfinance institution in Zimbabwe increased the validity of the information gathered and used. The precision of the instruments utilized is addressed by reliability. Snapshots of data and outcomes are shown at the end of the project to ensure the research credibility. For this study, programming in R was used for a variety of tasks. Scikit-learn was utilized for the machine learning models.

### **3.7 Chapter conclusion**

This chapter expanded on the data collection process, the modelling process and the model evaluation techniques used in the data set. The following chapter covers the data presentation, analysis and interpretation.

## **CHAPTER 4: DATA PRESENTATION, ANALYSIS AND DISCUSSION**

### **4.0 Introduction**

This chapter focuses on data analysis and presentation of findings for modelling credit default risk in microfinance in Zimbabwe. The data analysis process involves exploring the data, cleaning and transforming the data, selecting features, and building machine learning models. It also discusses the procedures the researcher used to analyze the data. Additionally, it offers the outcomes of the investigations of the researcher into the machine learning models.

### **4.1 Data exploration and visualizations**

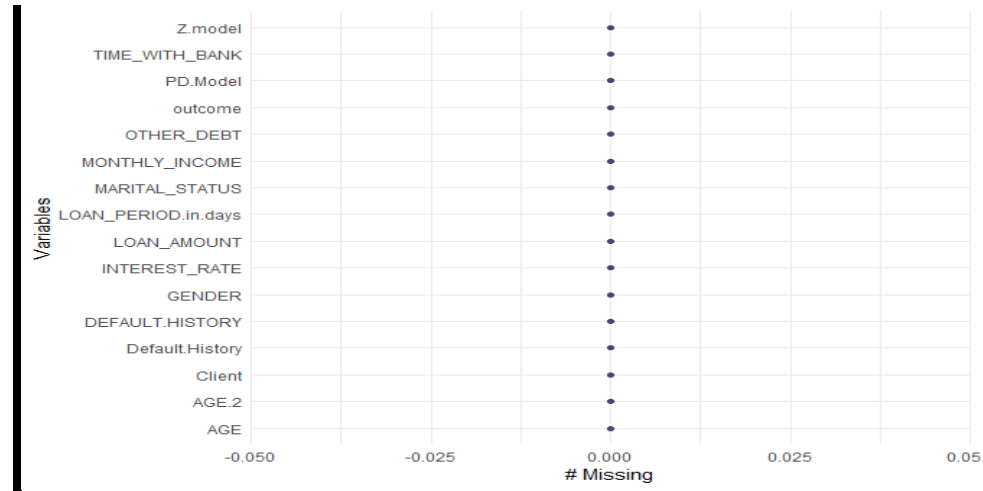
Data exploration and visualizations are important for modelling credit default risk in microfinance in Zimbabwe because they help to identify patterns and relationships in the data that can inform the development of machine learning models. By exploring the data and visualizing it in different ways, we can gain insights into the factors that contribute to credit default and the effectiveness of different features in predicting default risk. The reasons why data exploration and visualizations are important for modelling credit default risk in microfinance in Zimbabwe are explained and illustrated below.

#### **4.1.1 Identifying missing values**

Data exploration can help to identify missing values in the data, which need to be handled before building machine learning models. The `gg_miss_var()` function from the `naniar` package was used to visualize missing values in the dataset. This function creates bar charts for each variable in the dataset, displaying the percentage of missing values in each variable.

In the code above, the researcher loaded the credit data from a CSV file and use `gg_miss_var()` to visualize missing values in the dataset. This code requires the `naniar`

package to be installed. The plot was as follows, indicating that there were no missing values in all the variables which implies greater quality of the data.



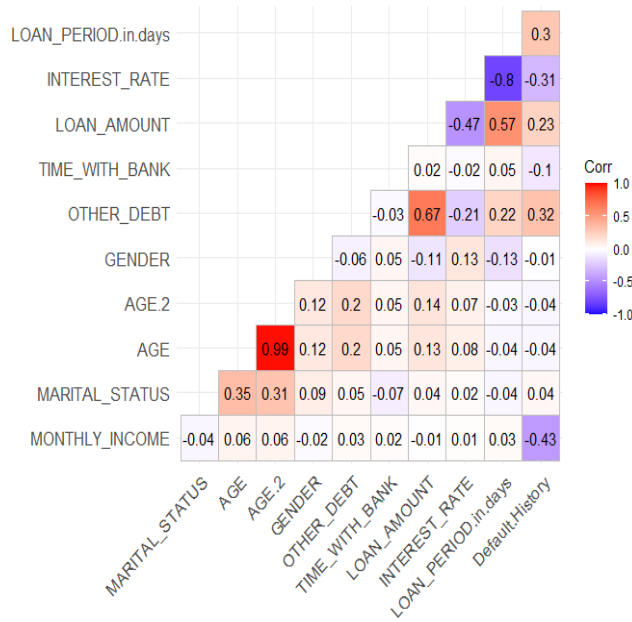
**Figure 4.1: Missing values**

#### **4.1.2 Understanding the distribution of the target variable:**

Visualizing the distribution of the target variable can help to understand the prevalence of credit default in the data and inform the choice of evaluation metrics for the machine learning models.

#### **4.1.3 Identifying correlations between features**

Exploring the correlation between variables can help to identify unnecessary or highly correlated features that may need to be removed before building machine learning models. The correlation matrix was plotted and depicted in the diagram as follows;



**Figure 4.2 : Correlation matrix**

From the matrix age and age squared are highly positively correlated so we remove variable age squared in the dataset. There is a negative correlation between loan period and interest rate and this assure us that the data is indeed legit by showing us trends and patterns which are already in credit data.

### 4.3 Models for credit default risk in microfinance

Credit default risk is the risk that an applicant may fail to repay a loan or credit debt. In microfinance, where lending is typically targeted towards low-income individuals or small businesses, credit default risk is a major concern for lenders. The following models were fitted.

#### 4.3.1 Stepwise Logistic model

The researcher first loaded the necessary libraries (tidyverse, caret, pROC) and the data was already loaded from the manipulations which were done before, then convert the categorical variables to factors and split the data into training and testing using the createDataPartition() function from the caret package.

The stepwise logistic regression model was build using the step() and glm() functions from the statistical package, and make predictions on both the training and testing data using the predict() function. The results from the model were as follows;

```
Call:
glm(formula = Default.History ~ MONTHLY_INCOME + LOAN_AMOUNT,
     family = "binomial", data = train_data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	8.2523	24.4082	0.338	0.735
MONTHLY_INCOME	-1.0438	0.8714	-1.198	0.231
LOAN_AMOUNT	0.5209	0.4342	1.200	0.230

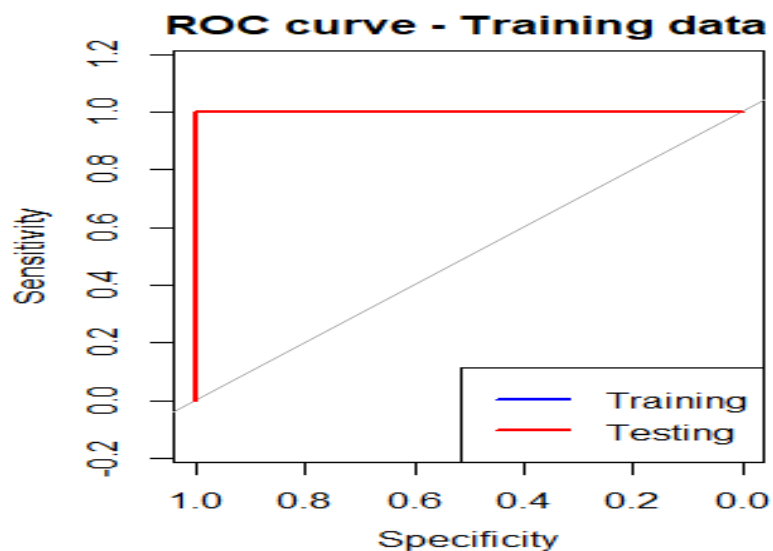
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 5.0294e+04 on 37359 degrees of freedom  
Residual deviance: 9.8085e-03 on 37357 degrees of freedom  
AIC: 6.0098

Number of Fisher Scoring iterations: 25

The model's performance was assessed using quality metrics.

such as F1 score, AUC and visualizations like ROC curves and histograms. The ggplot() function from the ggplot2 package to create the histograms.



**Figure 4.3: ROC curve for stepwise logistic regression**

Finally, the performance of the model was evaluated using performance metrics such as AUC (area under the ROC curve) for both the training and testing data using the roc() function from the pROC package.

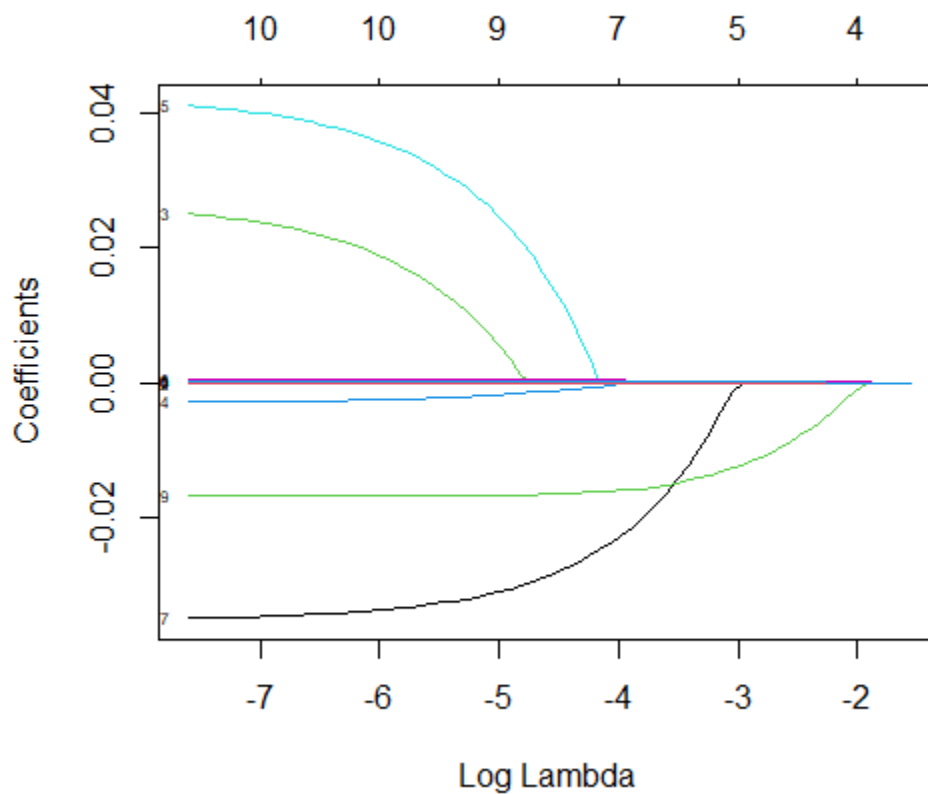
```
> print(paste0("AUC (Training): ", round(auc_train, 2)))  
[1] "AUC (Training): 1"  
> print(paste0("AUC (Testing): ", round(auc_test, 2)))  
[1] "AUC (Testing): 1"
```

An AUC of 1 for both the training and testing sets indicates that the model has a perfect ability to distinguish between positive and negative outcomes. This means that the model is able to correctly classify all positive and negative instances without making any errors.

### **4.3.2 Least Absolute Shrinkage and Selection Operator (LASSO)**

The necessary libraries (tidyverse, glmnet, pROC) were loaded and the data (default\_risk\_data.csv), then converted the categorical variables to factors and split the data into training and testing sets using the createDataPartition() function from the caret package.

The LASSO model was fitted using the glmnet() function from the glmnet package and plot the coefficient paths using the plot() function and choose the optimal value of lambda using cross-validation with the cv.glmnet() function. The path is as follows;

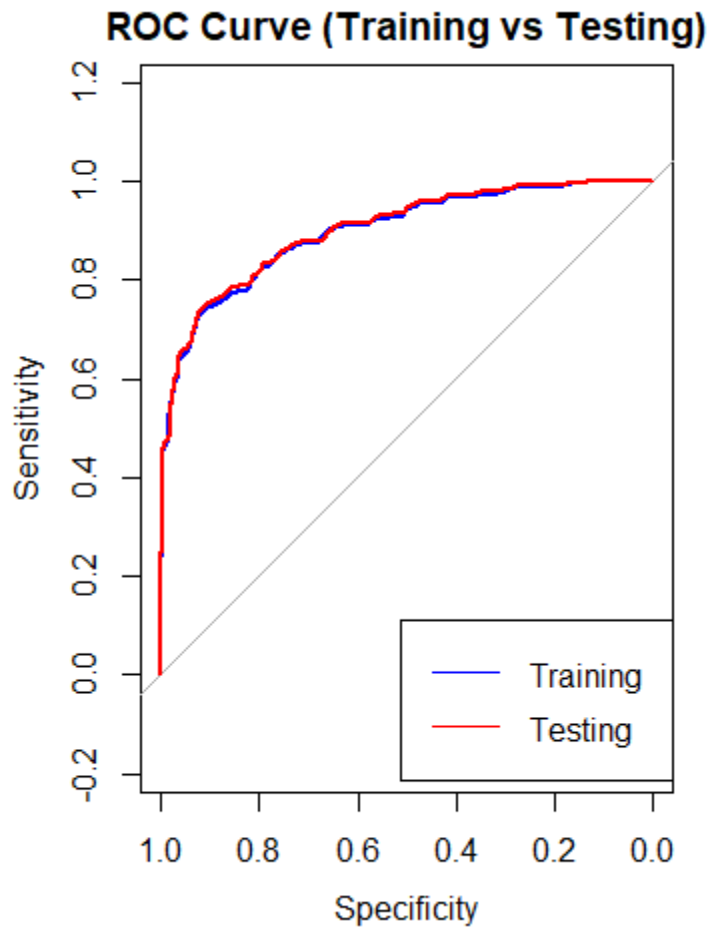


**Figure 4.4: LASSO coefficients path**

Model coefficients were as follows;

```
> coef(fit, s = lambda_opt)
11 x 1 sparse Matrix of class "dgCMatrix"
      s1
(Intercept)      1.077763e+00
Client           -2.125272e-07
MONTHLY_INCOME  -2.508096e-04
MARITAL_STATUS1  2.501130e-02
AGE              -2.934594e-03
GENDER1          4.102844e-02
OTHER_DEBT       5.054733e-04
TIME_WITH_BANK  -3.499812e-02
LOAN_AMOUNT      -3.379056e-06
INTEREST_RATE    -1.688243e-02
LOAN_PERIOD. in. days  9.555486e-05
```

Predictions are made on the training and testing data using the predict() function and calculate the ROC curves using the roc() function from the pROC package. We plot the ROC curves using the plot() function and calculate the AUC values using the roc() function. The AUC value is 0.898



**Figure 4.5: LASSO ROC curve**

### 4.3.3 Neural Network- multi-layer perceptron

Firstly the necessary libraries (tidyverse, nnet, caret) were loaded, then converted the categorical variables to factors and split the data into training and testing sets using the `createDataPartition()` function from the caret package.

Multi-layer perceptron neural network was fitted using the `nnet()` function from the nnet package. A hidden layer size of 5 and a weight decay of  $1e-5$  was specified. The multi-layer perceptron was fitted as follows;

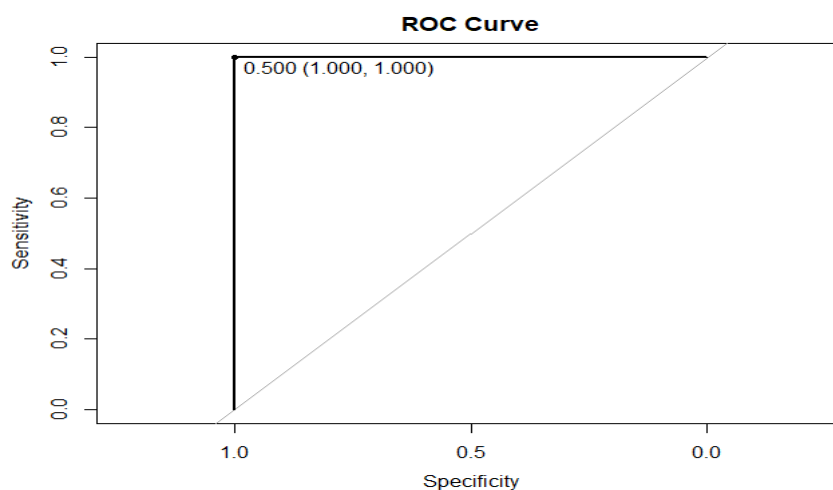


```

> mlp_fit <- nnet(Default.History ~ ., data = train_data, size = 5, decay = 1e-5)
# weights: 61
initial value 27720.409101
iter 10 value 16374.389362
iter 20 value 1423.950735
iter 30 value 1157.937288
iter 40 value 787.263166
iter 50 value 255.880656
iter 60 value 226.524199
iter 70 value 223.366186
iter 80 value 121.618998
iter 90 value 7.832452
iter 100 value 5.108192
final value 5.108192
stopped after 100 iterations

```

Predictions on the training and testing data using the `predict()` function with the `type = "raw"` argument were made. Finally, we print the performance metrics and plot the ROC curve for the testing data using the `roc()` function from the `pROC` package. The AUC value is 0.997



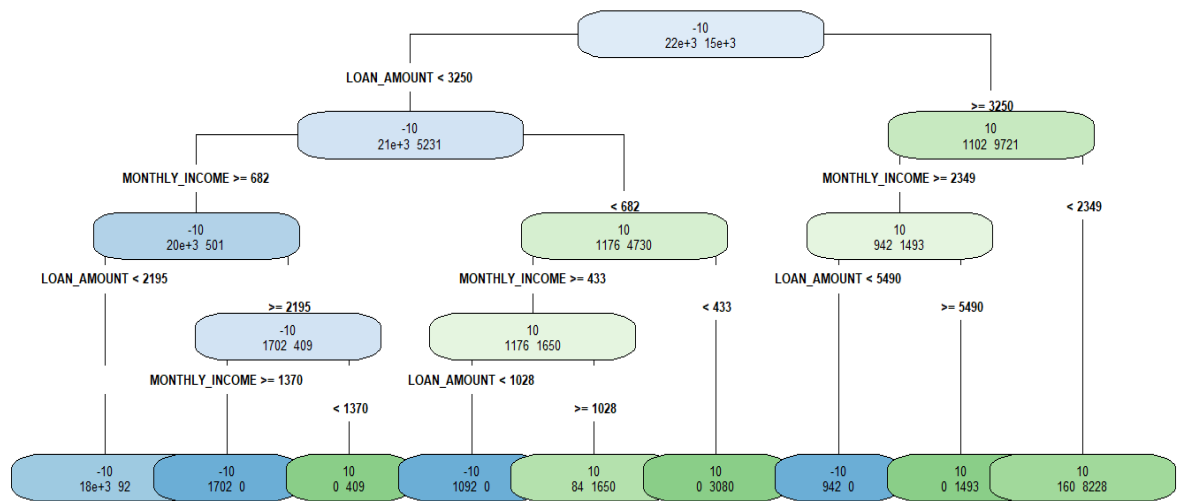
**Figure 4.6: NN ROC curve**

#### 4.3.4 Decision trees

The necessary libraries (`tidyverse`, `rpart`, `rpart.plot`) were loaded, then convert the categorical variables to factors and split the data into training and testing sets using the `createDataPartition()` function from the `caret` package.

The researcher then fit a decision tree using the `rpart()` function from the `rpart` package and specify the `method = "class"` argument to perform classification. The decision tree

was plotted using the `rpart.plot()` function from the `rpart.plot` package. We specify the `type = 4` argument to display the proportion of observations in each node, and the `extra = 1` argument to display the misclassification rate in each node.



**Figure 4.7: Decision tree**

The researcher made predictions on the training and testing data using the `predict()` function with the `type = "class"` argument, then calculate quality metrics (accuracy and F1 score) using the `confusionMatrix()` function from the `caret` package. Finally, we print the performance metrics and plot the decision tree using the `rpart.plot()` function.

```
> # Print performance metrics
> cat("Training Accuracy:", train_accuracy, "\n")
Training Accuracy: 0.9910064
> cat("Training F1 Score:", train_f1_score, "\n")
Training F1 Score: 0.9924772
>
> cat("Testing Accuracy:", test_accuracy, "\n")
Testing Accuracy: 0.9913276
> cat("Testing F1 Score:", test_f1_score, "\n")
Testing F1 Score: 0.9927439
```

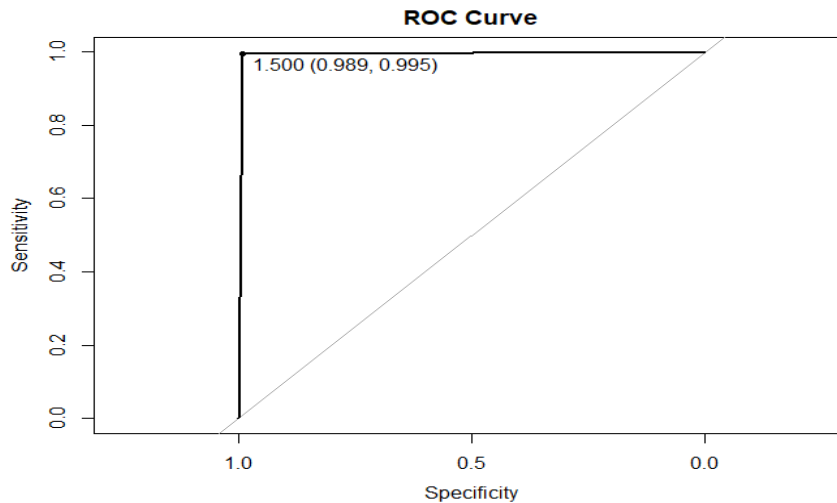
The researcher added the `roc()` function from the `pROC` package to calculate the area under the ROC curve (AUC) for the training and testing data and passed the true class labels for each dataset (`train_data$Default.History` and `test_data$Default.History`) and the predicted class probabilities (`as.numeric(train_pred)` and `as.numeric(test_pred)`) as arguments to the function, then print the AUC scores for each dataset using the `cat()` function.

```

> # Print performance metrics
> cat("Training AUC:", train_auc, "\n")
Training AUC: 0.991479
> cat("Testing AUC:", test_auc, "\n")
Testing AUC: 0.9918803

```

Finally, the ROC curve was plotted for the testing data using the plot() function with the test\_roc object as its argument. The resulting plot shows the ROC curve along with the optimal threshold value.

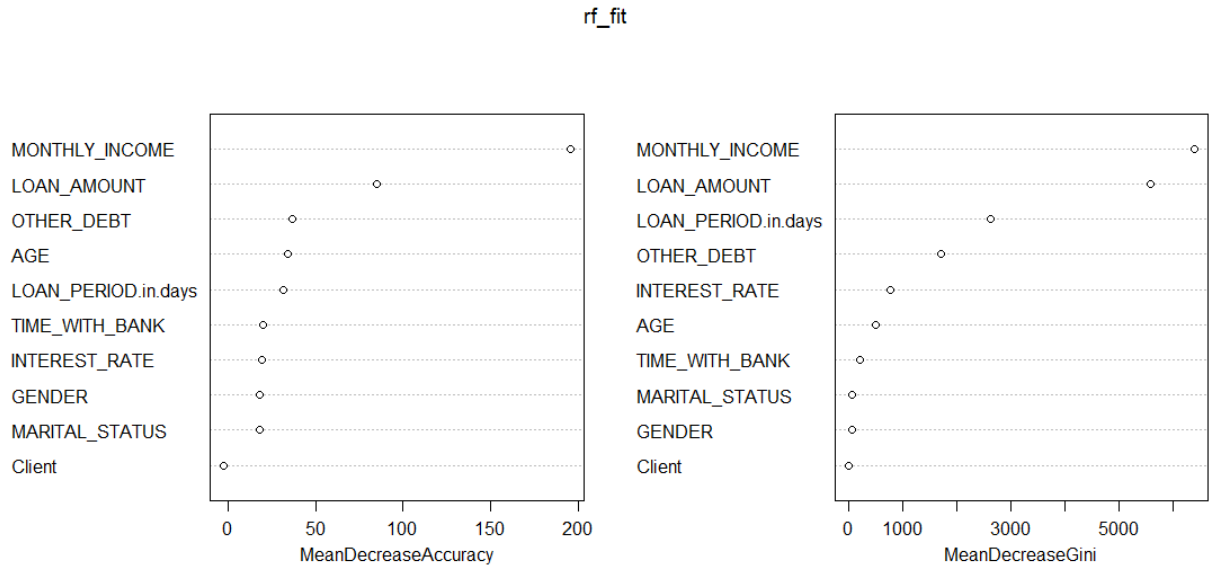


**Figure 4.8: Decision tree ROC curve**

### 4.3.5 Random forests

The researcher loaded the necessary libraries (tidyverse, randomForest, caret, pROC) and the data (credit\_default\_data.csv), then converted the categorical variables to factors and split the data into training and testing sets using the createDataPartition() function from the caret package.

The researcher fitted a random forest using the randomForest() function from the random Forest package and then specified the ntree = 500 argument to build 500 trees, and the importance = TRUE argument to calculate variable importance. the variable importance was plotted using the varImpPlot() function from the randomForest package and the output was as follows;



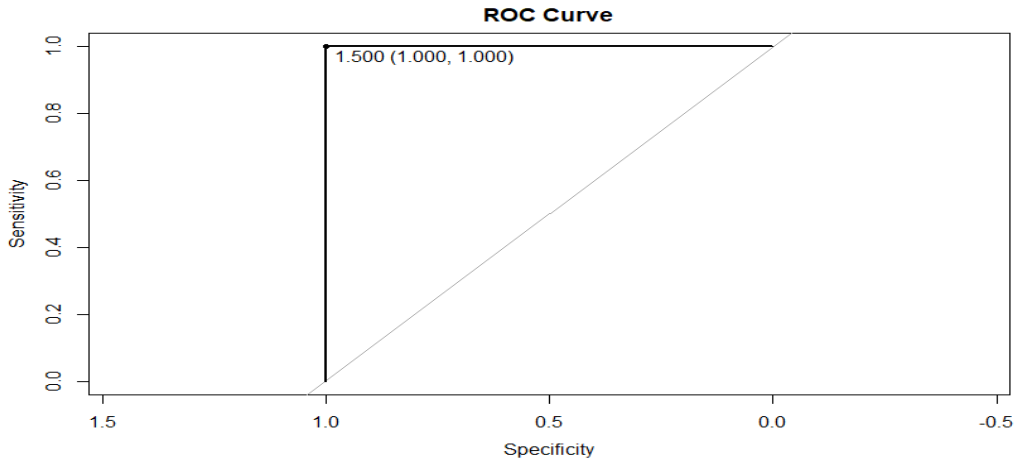
**Figure 4.9: Variable importance plot**

Predictions on the training and testing data using the predict() function were made, then calculated quality metrics (accuracy and F1 score) using the confusionMatrix() function from the caret package.

```
> # Print performance metrics
> cat("Training Accuracy:", train_accuracy, "\n")
Training Accuracy: 1
> cat("Training F1 Score:", train_f1_score, "\n")
Training F1 Score: 1
>
> cat("Testing Accuracy:", test_accuracy, "\n")
Testing Accuracy: 1
> cat("Testing F1 Score:", test_f1_score, "\n")
Testing F1 Score: 1
```

Finally, the performance metrics were printed, calculated the AUC score and plot the ROC curve for the testing data using the roc() and plot() functions from the pROC package.

```
> # Print AUC score
> cat("Testing AUC:", test_auc, "\n")
Testing AUC: 1
```



**Figure 4.10: RF ROC curve**

#### 4.4. Model analysis and discussion

In this chapter, five different models are explored for credit default modelling in microfinance in Zimbabwe which are stepwise logistic regression, LASSO model, neural networks, decision trees, and random forest. The stepwise logistic regression and LASSO model provided a baseline for comparison with the more complex models. This agrees with the results in a research by Haro and Germano et.al (2019). These models were able to identify significant predictors of credit default, but their predictive performance was limited compared to the tree-based models. The neural network model provided a flexible and powerful way to model credit default, but it was challenging to interpret and required more computational resources to train.

The decision tree model provided a transparent and interpretable way to model credit default, but its performance was limited compared to the random forest model. The random forest model used cast of characters of decision trees to improve predictive performance and had the highest testing performance and AUC score among all the models. This finding is in line with the results obtained by Coser (2019).

**Table 4.1: Summary of the performance evaluation results of the models**

Model	Precision score	Recall score	F1 score	Accuracy	AUC
Decision	1	1	0.993	0.991	0.992

tree					
Lasso	NA	NA	NA	NaN	0.898
Step wise	1	1	0.94	1	1
Random forest	1	0.99	1	1	0.992
Neural network	NA	NA	NA	0.022	0.997

The performance evaluation findings of the ML models fitted in this research are shown in Table 4.1. The precision score, recall score, F1-score, Accuracy, and AUC values were among the performance evaluation measures. The F1 score is a machine learning evaluation statistic that combines precision and recall values to measure a model's accuracy. Precision measures how many of the model's positive predictions were correct, whereas recall measures how many of the dataset's positive class samples were correctly identified by the model. The AUC value summarizes the overall diagnostic accuracy of the model. The result NA indicates that the dataset was not suitable for that particular machine learning model, resulting in model failure.

Stepwise and random forest has the highest accuracy score followed by decision tree and then Neural network and Lasso. On the other hand the recall score for the decision tree and step wise were highest followed by random forest.

Random forest has the highest F1 score, followed by decision tree and stepwise, and lasso has the lowest. Except for Lasso, all of the models got relatively high accuracy scores. A high accuracy score indicates that the models can capture the defaults. The AUC for all of the models was greater than 0.5, indicating that all of the models did well in catching false negatives and positives, indicating that they are all skilled models, and there were enough characteristics for the models to train well, Dumitrescu et.al (2022).

The study optimized the two top performing models, random forest and decision trees, based on the F1 score and AUC value. Finally, the random forest model is the best model. The findings in this study corroborate with the results in a research carried by Coser (2019) and Madaan et.al (2021).

#### **4.5 Chapter conclusion**

In this chapter, the researcher explored five different models for credit default modelling in microfinance in Zimbabwe which are stepwise logistic regression, LASSO model, neural networks, decision trees, and random forests. Random forest model is the most efficient model for modelling credit default risk. The next chapter is providing detailed conclusions and recommendations supported on the findings from this chapter.

## **CHAPTER 5: SUMMARY, CONCLUSION AND RECOMMENDATIONS**

### **5.1 Introduction**

This chapter gives the final conclusions of the main findings of the research. It also gives the summary of the project. It also provide various recommendations to the microfinance institutions in Zimbabwe.

### **5.2 Summary of the findings**

From the data utilized in this study, it was discovered that machine learning models performed better in modelling credit default risk in MFIs. The study found that the stepwise logistic regression and LASSO model were able to identify significant predictors of credit default, but their predictive performance was limited compared to the tree-based models. The neural network model provided a flexible and powerful way to model credit default, but it was challenging to interpret and required more computational resources to train.

The decision tree model provided a transparent and interpretable way to model credit default, but its performance was limited compared to the random forest model. The random forest model used an ensemble of decision trees to improve predictive performance and had the highest testing performance and AUC score among all the models. Therefore, the random forest model is the best model for credit default modelling in microfinance in Zimbabwe.

The study also found that borrower characteristics, such as age, gender, income, and education level, are significant predictors of credit default.

### **5.3 Conclusions**

This study concluded that machine learning models are better in predicting the



defaults of loans. Therefore, MFIs in Zimbabwe can implement the machine learning model in their operation as a way of reducing credit risks at the organisation. The implementation of machine learning in modelling credit risk require large amount of data and more features should be acquired by the MFIs before the loans are issued to clients to ensure their long-term sustainability.

#### **5.4 Recommendations**

Several recommendations for microfinance institutions in Zimbabwe can be made based on the results of the research.

1. It is recommended that microfinance institutions use the random forest model for credit default modelling. This model provides the highest predictive performance and AUC score and can help institutions identify potential defaulters and manage credit risk effectively.

2. It is recommended that microfinance institutions collect more detailed borrower information to improve credit default modelling. This could include data on social, economic, and demographic factors that may impact creditworthiness. By collecting more detailed borrower information, institutions can improve the accuracy of credit default models and make better lending decisions.

3. It is recommended that microfinance institutions continually monitor and update their credit default models to adapt to changing market conditions and borrower behaviour. As the industry evolves, institutions must remain vigilant and proactive in managing credit risk to ensure their long-term sustainability.

#### **5.5 Chapter conclusion**

The research findings were summarised in this chapter. Additionally, conclusions and recommendations were also provided.

## REFERENCES

- Abafita, J. (2003). Microfinance and loan repayment performance. *A Case Study of the Oromia Credit and Savings Share*
- AL-SHAYEA, Q. K., & EL-REFAE, G. A. (2011). Evaluating credit risk using artificial neural networks. *Global Engineers & Technologist Review*, 1(1).
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), 589-609.
- Amir, M. K. B. A., & Choudhury, N. N. (2023). The impact of non-performing loans on bank lending behavior before and amid COVID-19 Pandemic: Evidence from selected private commercial banks in Bangladesh. *International Journal of Research in Business and Social Science (2147-4478)*, 12(3), 272-285.
- Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press.
- Arene, C. J. (1993). An analysis of loan repayment potentials of smallholder soya bean group farmers in Nigeria. *Quarterly journal of international agriculture (Germany)*.
- Awunyo-Vitor, D. (2012). Determinants of loan repayment default among farmers in Ghana.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.
- Bland, J. M., & Altman, D. G. (1994). Statistics notes: some examples of regression towards the mean. *Bmj*, 309(6957), 780.
- Breiman, L. (1996). Bagging predictors. *Machine learning*, 24, 123-140.
- Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
- Brynjolfsson, E., & McAfee, A. N. D. R. E. W. (2017). Artificial intelligence, for real. *Harvard business review*, 1, 1-31.

- Carling, K., Jacobson, T., Lindé, J., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of banking & finance*, 31(3), 845-868.
- Casu, B., & Girardone, C. (2006). Bank competition, concentration and efficiency in the single European market. *The Manchester School*, 74(4), 441-468.
- Chaves, R. A., & González Vega, C. (1992). Principles of Regulation and Prudential Supervision: Should They Be Different for Microenterprise Finance Organizations?.
- Coravos, A. R. (2010). *Measuring the Likelihood of Small Business Loan Default* (Doctoral dissertation, Duke University).
- Coşer, A., Maer-matei, M. M., & Albu, C. (2019). PREDICTIVE MODELS FOR LOAN DEFAULT RISK ASSESSMENT. *Economic Computation & Economic Cybernetics Studies & Research*, 53(2).
- Crosbie, P., & Bohn, J. (2019). Modeling default risk. In *World Scientific Reference on Contingent Claims Analysis in Corporate Finance: Volume 2: Corporate Debt Valuation with CCA* (pp. 471-506).
- Csizmadia, A., Curzon, P., Dorling, M., Humphreys, S., Ng, T., Selby, C., & Woollard, J. (2015). Computational thinking-A guide for teachers.
- Dlamini, B., & Mbira, L. (2017). The current Zimbabwean liquidity crisis: a review of its precipitates. *Journal of Economics and Behavioral Studies*, 9(3 (J)), 212-219.
- Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178-1192.
- Duy, V. Q., Loc, B. T., Ngoc, N. P. H., Ngon, Q. L., Ha, T. T., Thi, D., & Lieu, T. CREDIT RISK IN GROUP LENDING MODEL AT THE BANK FOR SOCIAL POLICIES IN HAU GIANG PROVINCE.
- El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* (pp. 3-11). Springer International Publishing.

- Erickson, M. L., Elliott, S. M., Brown, C. J., Stackelberg, P. E., Ransom, K. M., Reddy, J. E., & Cravotta III, C. A. (2021). Machine-learning predictions of high arsenic and high manganese at drinking water depths of the glacial aquifer system, northern continental United States. *Environmental Science & Technology*, 55(9), 5791-5805.
- Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2), 368-378.
- Fisher, R. A. (1936). Statistical methods for research workers. *Statistical methods for research workers.*, (6th Ed).
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1), 1-58.
- Grigoli, F., & Mota, J. M. (2017). Interest rate pass-through in the Dominican Republic. *Latin American Economic Review*, 26(1), 4.
- Hao, X. (2014). *Credit Scoring Method and System Development for Imbalanced Datasets* (Doctoral dissertation, Fukushima University).
- Haro Alonso, D., Wernick, M. N., Yang, Y., Germano, G., Berman, D. S., & Slomka, P. (2019). Prediction of cardiac death after adenosine myocardial perfusion SPECT based on machine learning. *Journal of Nuclear Cardiology*, 26, 1746-1754.
- Haron, H., Said, S. B., Jayaraman, K., & Ismail, I. (2013). Factors influencing small medium enterprises (SMES) in obtaining loan. *International Journal of Business and Social Science*, 4(15).
- Hilbe, J. M. (2009). *Logistic regression models*. CRC press.
- Ibekwe, U. C., & Ukoha, I. I. (2011). Determinants of loan acquisition from the financial institutions by small-scale farmers in Ohafia Agricultural zone of Abia State, South East Nigeria. *Journal of Development and Agricultural Economics*, 3(2), 69-74.
- Jarrow, R. A., & Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. *The journal of finance*, 50(1), 53-85.

Klein, T. M. (1994). *External debt management: an introduction* (Vol. 23). World Bank Publications.

Kohansal, M. R., & Mansoori, H. (2009, October). Factors affecting on loan repayment performance of farmers in Khorasan-Razavi province of Iran. In *Conference on International Research on Food Security, Natural Resource Management and Rural Development, University of Hamburg* (Vol. 26, pp. 359-366).

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012042). IOP Publishing.

Malik, M., & Thomas, L. C. (2010). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society*, 61(3), 411-420.

MANGUDYA, D. (2016). Monetary Policy Statement. Retrieved March, 25, 2016.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. *Journal of banking & finance*, 1(3), 249-276.

Meurer, W. J., & Tolles, J. (2017). Logistic regression diagnostics: understanding how well a model predicts outcomes. *Jama*, 317(10), 1068-1069.

Mokhtar, S. H., Nartea, G., & Gan, C. (2012). Determinants of microcredit loans repayment problem among microfinance borrowers in Malaysia. *International Journal of Business and Social Research (IJBSR)*, 2(7), 33-45.

Muriithi, M. W. (2013). *The causes of non-performing loans in commercial banks in Kenya* (Doctoral dissertation, University of Nairobi).

Nyangena, B. O. (2019). *Consumer credit risk modelling using machine learning algorithms: a comparative approach* (Doctoral dissertation, Strathmore University).

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109-131.

- Orlova, E. V. (2021). Methodology and models for individuals' creditworthiness management using digital footprint data and machine learning methods. *Mathematics*, 9(15), 1820.
- Roodman, D. (2012). *Due diligence: An impertinent inquiry into microfinance*. CGD Books.
- Salas, V., & Saurina, J. (2002). Credit risk in two institutional regimes: Spanish commercial and savings banks. *Journal of Financial Services Research*, 22(3), 203.
- Samreen, A., & Zaidi, F. B. (2012). Design and development of credit scoring model for the commercial banks of Pakistan: Forecasting creditworthiness of individual borrowers. *International Journal of Business and Social Science*, 3(17).
- Segal, T. (2020). Nonperforming Loan (NPL).
- Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. *Journal of thoracic disease*, 11(Suppl 4), S574.
- Tien Bui, D., Tuan, T. A., Hoang, N. D., Thanh, N. Q., Nguyen, D. B., Van Liem, N., & Pradhan, B. (2017). Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization. *Landslides*, 14, 447-458.
- Wang, C., Han, D., Liu, Q., & Luo, S. (2018). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. *IEEE Access*, 7, 2161-2168.
- Warue, B. N. (2013). The effects of bank specific and macroeconomic factors on nonperforming loans in commercial banks in Kenya: A comparative panel data analysis. *Advances in Management and Applied Economics*, 3(2), 135.
- Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, 15(3), 757-770.

Wongnaa, C. A., & Awunyo-Vitor, D. (2013). Factors affecting loan repayment performance among yam farmers in the Sene District, Ghana.

Zhang, Y., Gan, C., & Li, Z. (2012). Effects of borrowers' quality on the size of market response to bank loan announcements in China. *Management Research Review*, 35(5), 379-404.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, 59-82.

Zohair, M. (2013). Factors Affecting Repayment of Loans by Micro-borrowers in Tunisia: An Empirical Study. *Journal of Management and Public Policy*, 4(2), 4-16.

## APPENDICES

### APPENDIX A: Snapshot of the data

Client	MONTHLY_I	MARITAL_S	AGE	AGE.2	GENDER	OTHER_DEI	TIME_WITH	LOAN_AMO	INTEREST_I
1	2897	1	49	2401	1	556,67	2	1000	17
2	2175	0	32	1024	0	258,68	5	1150	18
3	127	1	41	1681	0	309,42	5	1360	18
4	2565	1	35	1225	1	197,92	3	2500	18
5	439	0	33	1089	0	0	2	800	18
6	1625	1	33	1089	1	116	5	1200	18
7	2101	1	56	3136	1	160,37	5	970	18
8	508	1	34	1156	1	403,91	5	1580	18
9	2017	1	36	1296	1	42,92	4	4000	17,25
10	2562	0	27	729	1	166,77	3	9280	17,25
11	2038	0	47	2209	0	444,17	3	1000	18
12	330	1	49	2401	1	308,81	2	28720	17,25
13	642	0	25	625	1	534,71	5	6030	18
14	948	1	41	1681	1	191,25	4	1100	18
15	507	1	46	2116	1	439,33	5	2070	18
16	385	1	50	2500	1	274,97	2	5000	18
17	365	0	31	961	1	177,69	5	987	18
18	1451	0	28	784	1	255,99	2	1300	18
19	2129	0	50	2500	0	157,63	5	600	18
20	153	1	33	1089	1	183,84	5	1000	18
21	1623	0	38	1444	1	346,63	5	970	18
22	1091	1	38	1444	1	142,2	5	1000	18
23	1912	1	34	1156	0	275,61	5	1290	18
24	864	1	47	2209	1	121,22	2	1350	18

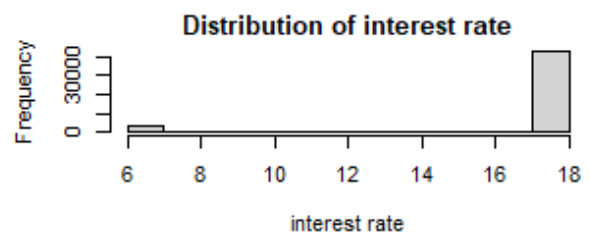
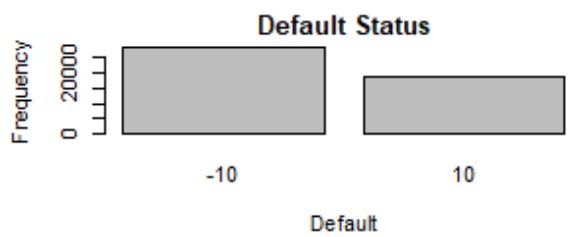
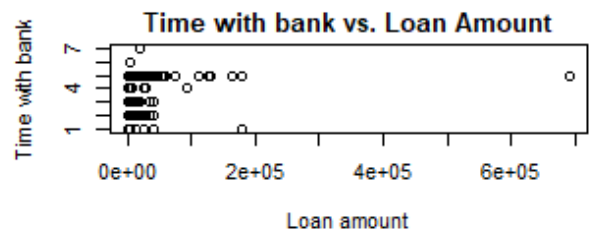
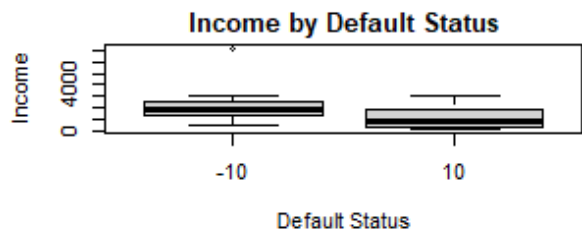
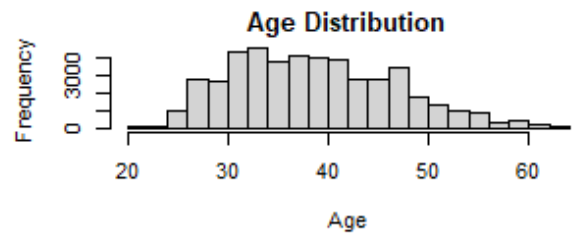
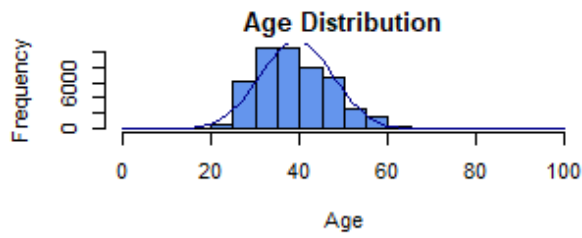


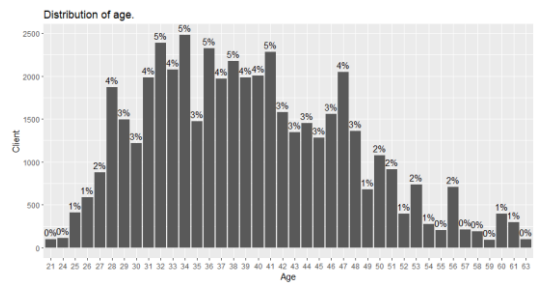
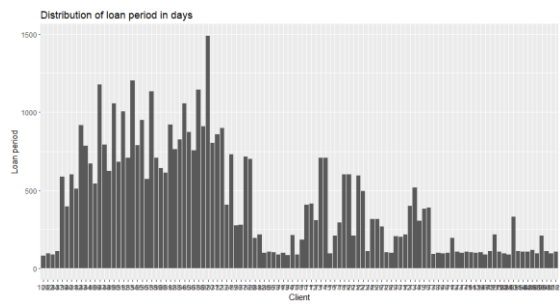
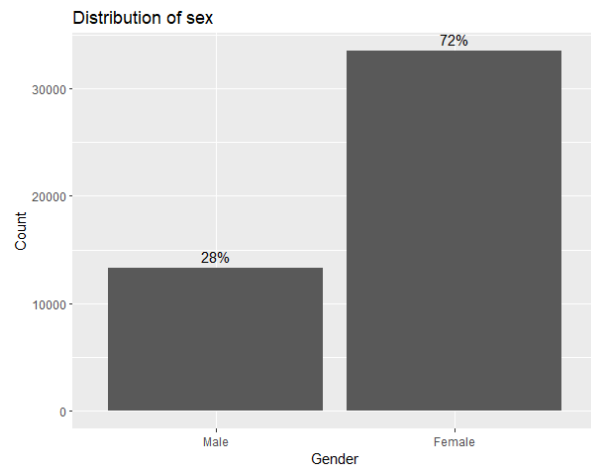
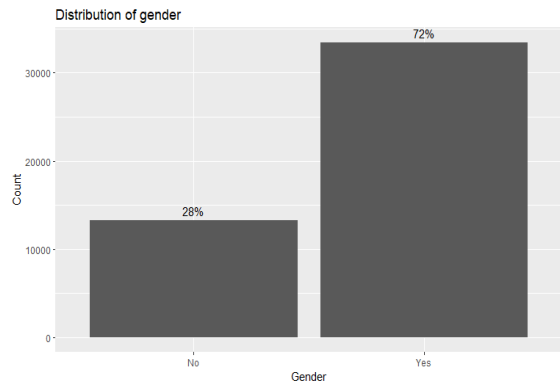
## APPENDIX B: Missing values and variable names

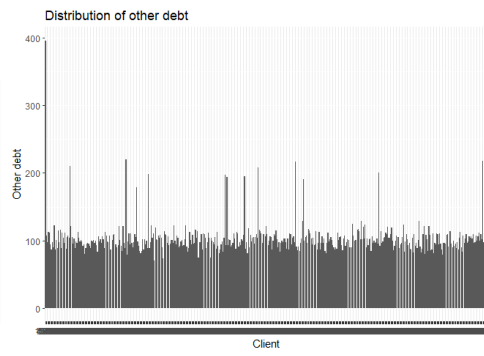
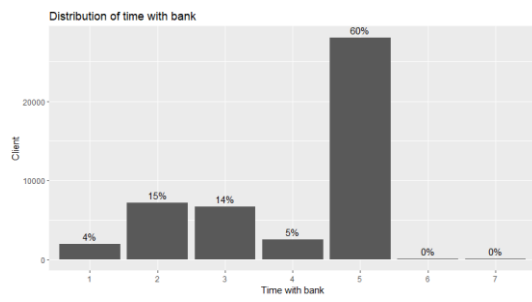
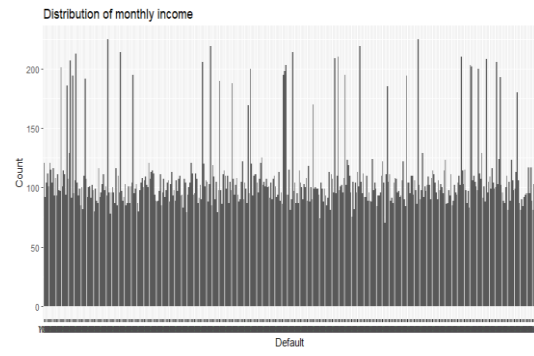
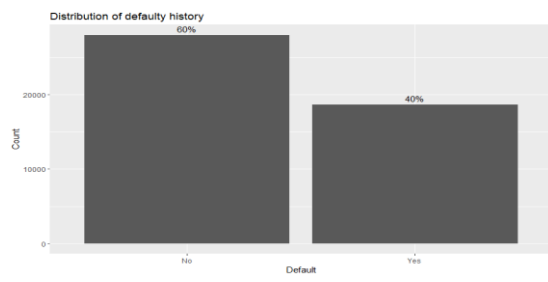
```
> View(credit_data)> # Check missing values> #install.packages("naniar")> library(naniar)>
print(colSums(is.na(df)))      Client  MONTHLY_INCOME  MARITAL_STATUS
      0          0          0
      AGE      AGE.2      GENDER
      0          0          0
      OTHER_DEBT  TIME_WITH_BANK  LOAN_AMOUNT
      0          0          0
      INTEREST_RATE LOAN_PERIOD.in.days  Default.History
      0          0          0 > gg_miss_var(credit_data)> # Check variable names>
names(df) [1] "Client"      "MONTHLY_INCOME"  "MARITAL_STATUS"
[4] "AGE"      "AGE.2"      "GENDER"
[7] "OTHER_DEBT"  "TIME_WITH_BANK"  "LOAN_AMOUNT"
[10] "INTEREST_RATE"  "LOAN_PERIOD.in.days" "Default.History"

>
```

## APPENDIX C: Distribution of variables







## APPENDIX D: R studio; R code

```
# Load libraries
library(tidyverse)

# Load data
df<- read_csv("credit_data.csv")
View(credit_data)

# Check missing values
#install.packages("naniar")
library(naniar)
print(colSums(is.na(credit_data)))
gg_miss_var(credit_data)

# Check variable names
names(credit_data)

# Drop variables
df<- select(credit_data, -outcome, -PD.Model, -Z.model,-DEFAULT.HISTORY)

# View updated dataset
head(df)
View(df)

# Convert all variables to numeric
df <- mutate_all(df, as.numeric)

#saving the truncated data
write_csv(df, "credit_data.csv")

# Visualize distribution of target variable
# Set graphics parameters
par(mfrow = c(3, 2), mar = c(4, 4, 2, 1), oma = c(0, 0, 2, 0))

# Create plots
hist(df$AGE, main = "Age Distribution", xlab = "Age", col = "cornflowerblue", breaks
= seq(0, 100, by = 5))
curve(dnorm(x, mean = mean(df$AGE), sd = sd(df$AGE)) * length(df$AGE) *
diff(seq(0, 100, by = 5))[1], add = TRUE, col = "darkblue")
hist(df$AGE, main = "Age Distribution", xlab = "Age")
boxplot(df$MONTHLY_INCOME ~ df$Default.History, main = "Income by Default
Status", xlab = "Default Status", ylab = "Income")
```

```

plot(df$LOAN_AMOUNT, df$TIME_WITH_BANK, main = "Time with bank vs.
Loan Amount", xlab = "Loan amount", ylab = "Time with bank")
barplot(table(df$Default.History), main = "Default Status", xlab = "Default", ylab =
"Frequency")
hist(df$INTEREST_RATE, main = "Distribution of interest rate", xlab = "interest rate")
dev.off()
ggplot(df, aes(x = factor(`MONTHLY_INCOME`))) +geom_bar() + labs(x = "Client",
y = "Income") +
  ggtitle("Distribution of monthly income")+scale_fill_manual(values = c("red",
"blue"))
ggplot(df, aes(x = factor(`AGE`))) +geom_bar() + labs(x = "Age", y = "Client") +
ggtitle("Distribution of age.")+
  scale_fill_manual(values = c("red", "green"))+
  geom_text(stat = "count", aes(label = paste0(round(..count../sum(..count..), digits = 2)
* 100, "%")), vjust = -0.5)
ggplot(df, aes(x = factor(`GENDER`))) +geom_bar() + labs(x = "Gender", y = "Count")
+
  scale_x_discrete(labels = c("Male", "Female")) + ggtitle("Distribution of sex") +
  scale_fill_manual(values = c("yellow", "orange"))+
  geom_text(stat = "count", aes(label = paste0(round(..count../sum(..count..), digits = 2)
* 100, "%")), vjust = -0.5)
ggplot(df, aes(x = factor(`OTHER_DEBT`))) +geom_bar() + labs(x = "Client", y =
"Other debt") +
  ggtitle("Distribution of other debt")+ scale_fill_manual(values = c("grey", "blue"))

ggplot(df, aes(x = factor(`TIME_WITH_BANK`))) +geom_bar() + labs(x = "Time
with bank", y = "Client") +
  ggtitle("Distribution of time with bank")+ scale_fill_manual(values = c("brown",
"blue"))+
  geom_text(stat = "count", aes(label = paste0(round(..count../sum(..count..), digits = 2)
* 100, "%")), vjust = -0.5)
ggplot(df, aes(x = factor(`LOAN_PERIOD.in.days`))) +geom_bar() + labs(x =
"Client", y = "Loan period") +

```

```

  ggtitle("Distribution of loan period in days")+ scale_fill_manual(values = c("red",
"purple"))
ggplot(df, aes(x = factor(`Default.History`))) +geom_bar() + labs(x = "Default", y =
"Count") +
  scale_x_discrete(labels = c("No", "Yes")) + ggtitle("Distribution of defaulty
history")+
  scale_fill_manual(values = c("red", "blue"))+
  geom_text(stat = "count", aes(label = paste0(round(..count../sum(..count..), digits = 2)
* 100, "%")), vjust = -0.5)
# Visualize correlation between features
# Calculate correlation matrix
cor_matrix <- cor(df[, c("MONTHLY_INCOME", "MARITAL_STATUS",
"AGE","AGE.2","GENDER" ,
"OTHER_DEBT","TIME_WITH_BANK", "LOAN_AMOUNT",
"INTEREST_RATE" ,
"LOAN_PERIOD.in.days", "Default.History")])
# Create correlation plot
library(ggcorrplot)
ggcorrplot(cor_matrix, type = "lower", lab = TRUE)

# Drop highly correlated variables
df<- select(df, -AGE.2)
names(df)

#Stepwise logistic model-----
# Load libraries
library(tidyverse)
library(caret)
library(pROC)

# Convert categorical variables to factors
df$Default.History <- as.factor(df$Default.History)
df$GENDER <- as.factor(df$GENDER)

```

```

df$MARITAL_STATUS <- as.factor(df$MARITAL_STATUS)
# Split data into training and testing sets
set.seed(123)
train_index <- createDataPartition(df$Default.History, p = 0.8, list = FALSE)
train_data <- df[train_index, ]
test_data <- df[-train_index, ]
# Build stepwise logistic model
logit_model <- step(glm(`Default.History` ~ ., data = train_data, family = "binomial"),
direction = "both")
# Make predictions on training and testing data
train_probs <- predict(logit_model, newdata = train_data, type = "response")
test_probs <- predict(logit_model, newdata = test_data, type = "response")
# Evaluate model performance using quality metrics
train_pred <- ifelse(train_probs > 0.5, "10", "-10")
test_pred <- ifelse(test_probs > 0.5, "10", "-10")
train_data$Default.History <- factor(train_data$Default.History, levels = c("10", "-10"))
train_pred <- factor(train_pred, levels = c("10", "-10"))
confusionMatrix(train_pred, train_data$Default.History)
test_data$Default.History <- factor(test_data$Default.History, levels = c("10", "-10"))
test_pred <- factor(test_pred, levels = c("10", "-10"))
confusionMatrix(test_pred, test_data$Default.History)
# Evaluate model performance using visualizations
plot(roc(train_data$Default.History, train_probs), col = "blue", main = "ROC curve -
Training data")
lines(roc(test_data$Default.History, test_probs), col = "red")
legend("bottomright", legend = c("Training", "Testing"), col = c("blue", "red"), lwd =
2)
ggplot(train_data, aes(x = `AGE`, fill = `Default.History`)) +
  geom_histogram(binwidth = 5, alpha = 0.5) +
  ggtitle("Age Distribution by Default Status - Training data") +
  xlab("Age") + ylab("Count") +
  scale_fill_manual(values = c("red", "blue"), name = "Default")
# Evaluate model performance using performance metrics

```



```

library(caret)
#install.packages("mltools")
library(mltools)
# Convert the predicted probabilities to class predictions using a threshold of 0.5
train_pred <- ifelse(train_probs > 0.5, "10", "-10")
# Calculate the F1 score
# Convert the predicted probabilities to class predictions using a threshold of 0.5
train_pred <- ifelse(train_probs > 0.5, "10", "-10")
# Calculate the F1 score
f1_train <- function(tp, fp, fn) {
  precision <- tp / (tp + fp)
  recall <- tp / (tp + fn)
  f1 <- 2 * precision * recall / (precision + recall)
  return(f1)
}
tp <- sum(train_pred == "10" & train_data$Default.History == "10")
fp <- sum(train_pred == "10" & train_data$Default.History == "-10")
fn <- sum(train_pred == "-10" & train_data$Default.History == "10")
f1_train <- f1_train(tp, fp, fn)
# Calculate the precision-recall AUC
pr_auc_train <- function(probs, actual, n = 1000) {
  # Sort the probabilities in descending order
  sorted <- sort(probs, decreasing = TRUE)

  n <- 1000 # number of thresholds to use
  prec <- numeric(n)
  rec <- numeric(n)
  for (i in 1:n) {
    threshold <- i / n # set the threshold to the ith quantile of the predicted probabilities
    preds <- ifelse(train_probs > threshold, "10", "-10") # set predicted labels based on
threshold
    tp <- sum(preds == "10" & train_data$Default.History == "10")
    fp <- sum(preds == "10" & train_data$Default.History == "-10")
    fn <- sum(preds == "-10" & train_data$Default.History == "10")

```

```

if (tp + fp == 0) {
  prec[i] <- 1 # if there are no predicted positives, set precision to 1
} else {
  prec[i] <- tp / (tp + fp)
}
if (tp + fn == 0) {
  rec[i] <- 1 # if there are no actual positives, set recall to 1
} else {
  rec[i] <- tp / (tp + fn)
}
print(paste0("AUC (Training): ", round(auc_train, 2)))
print(paste0("AUC (Testing): ", round(auc_test, 2)))
# Calculate the overall precision and recall
overall_tp <- sum(train_pred == "10" & train_data$Default.History == "10")
overall_fp <- sum(train_pred == "10" & train_data$Default.History == "-10")
overall_fn <- sum(train_pred == "-10" & train_data$Default.History == "10")
overall_prec <- overall_tp / (overall_tp + overall_fp)
overall_rec <- overall_tp / (overall_tp + overall_fn)
# Print the overall precision and recall
cat("Overall precision:", round(overall_prec, 2), "\n")
cat("Overall recall:", round(overall_rec, 2), "\n")

```

### **#Least Absolute Shrinkage and Selection Operator (LASSO)-----**

```

# Load libraries
library(tidyverse)
library(glmnet)
library(pROC)
# Fit LASSO model using glmnet
# Create model matrix and response vector
x_train <- model.matrix(`Default.History` ~ ., data = train_data)[, -1]
y_train <- as.numeric(train_data$Default.History) - 1
# Fit LASSO model using glmnet
fit <- glmnet(x_train, y_train, alpha = 1)
# Summarize the results
coef(fit, s = lambda_opt)

```

```

# Plot coefficient paths
plot(fit, xvar = "lambda", label = TRUE)
# Choose optimal lambda using cross-validation
cv_fit <- cv.glmnet(x_train, y_train, alpha = 1)
lambda_opt <- cv_fit$lambda.min
# Make predictions on training and testing data
x_test <- model.matrix(`Default.History` ~ ., data = test_data)[, -1]
y_test <- as.numeric(test_data$Default.History) - 1
train_pred <- predict(fit, s = lambda_opt, newx = x_train, type = "response")
test_pred <- predict(fit, s = lambda_opt, newx = x_test, type = "response")

# Calculate and plot ROC curves
train_auc <- roc(as.numeric(y_train), as.numeric(train_pred))
test_auc <- roc(as.numeric(y_test), as.numeric(test_pred))

plot(train_auc, col = "blue", main = "ROC Curve - Train vs Test", print.auc = TRUE)
lines(test_auc, col = "red", print.auc = TRUE, lty = 2)
legend("bottomright", c("Train", "Test"), col = c("blue", "red"), lty = c(1, 2))
# Calculate quality metrics
# Convert train_pred[,2] and y_train to factors with levels c(10, -10)
# Convert train_pred to factor with levels c(10, -10)
train_pred_factor <- factor(ifelse(train_pred > 0.5, 10, -10), levels = c(10, -10))
# Convert y_train to factor with levels c(10, -10)
y_train_factor <- factor(y_train, levels = c(10, -10))
# Calculate confusion matrix for training data
train_confusion <- confusionMatrix(train_pred_factor, y_train_factor)
# Print confusion matrix
print(train_confusion)
# Convert test_pred to factor with levels c(10, -10)
test_pred_factor <- factor(ifelse(test_pred > 0.5, 10, -10), levels = c(10, -10))
# Convert y_test to factor with levels c(10, -10)
y_test_factor <- factor(y_test, levels = c(10, -10))
# Calculate confusion matrix for test data
test_confusion <- confusionMatrix(test_pred_factor, y_test_factor)

```

```

# Print confusion matrix
print(test_confusion)
train_accuracy <- train_confusion$overall["Accuracy"]
test_accuracy <- test_confusion$overall["Accuracy"]
train_precision <- train_confusion$byClass["Pos Pred Value"]
test_precision <- test_confusion$byClass["Pos Pred Value"]
train_recall <- train_confusion$byClass["Sensitivity"]
test_recall <- test_confusion$byClass["Sensitivity"]
train_f1_score <- train_confusion$byClass["F1"]
test_f1_score <- test_confusion$byClass["F1"]
# Print performance metrics
cat("Training Accuracy:", train_accuracy, "\n")
cat("Testing Accuracy:", test_accuracy, "\n")
cat("Training Precision:", train_precision, "\n")
cat("Testing Precision:", test_precision, "\n")
cat("Training Recall:", train_recall, "\n")
cat("Testing Recall:", test_recall, "\n")
cat("Training F1 Score:", train_f1_score, "\n")
cat("Testing F1 Score:", test_f1_score, "\n")

#Neural Network- multi-layer perceptron-----
# Load libraries
library(tidyverse)
library(nnet)
library(caret)
# Convert categorical variables to factors
df$Default.History <- as.factor(df$Default.History)
df$GENDER <- as.factor(df$GENDER)
df$MARITAL_STATUS <- as.factor(df$MARITAL_STATUS)
# Fit multi-layer perceptron neural network
mlp_fit <- nnet(Default.History ~ ., data = train_data, size = 5, decay = 1e-5)
# Make predictions on training and testing data
train_pred <- predict(mlp_fit, newdata = train_data, type = "raw")
test_pred <- predict(mlp_fit, newdata = test_data, type = "raw")

```

```

# Calculate quality metrics
# Convert train_pred to factor with levels c(10, -10)
train_pred_factor <- factor(ifelse(train_pred > 0.5, 10, -10), levels = c(10, -10))
# Convert train_data$Default.History to factor with levels c(10, -10)
y_train_factor <- factor(train_data$Default.History, levels = c(10, -10))
# Calculate confusion matrix for training data
train_confusion_neural_network <- confusionMatrix(train_pred_factor, y_train_factor)
# Print confusion matrix
print(train_confusion_neural_network)
# Convert test_pred to factor with levels c(0, 1)
test_pred_factor <- factor(ifelse(test_pred > 0.5, 10, -10), levels = c(10, -10))
# Convert test_data$Default.History to factor with levels c(10, -10)
y_test_factor <- factor(test_data$Default.History, levels = c(10, -10))
# Calculate confusion matrix for test data
test_confusion_neural_network <- confusionMatrix(test_pred_factor, y_test_factor)
# Print confusion matrix
print(test_confusion_neural_network)
# Print performance metrics
cat("Training Accuracy:", train_accuracy, "\n")
cat("Training F1 Score:", train_f1_score, "\n")
cat("Testing Accuracy:", test_accuracy, "\n")
cat("Testing F1 Score:", test_f1_score, "\n")
# Plot ROC curve and confusion matrix for testing data
test_roc <- roc(test_data$Default.History, test_pred)
plot(test_roc, print.thres = "best", main = "ROC Curve")

#Decision trees-----
# Load libraries
library(tidyverse)
library(rpart)
library(rpart.plot)
# Fit decision tree
tree_fit <- rpart(Default.History ~ ., data = train_data, method = "class")
# Plot decision tree

```

```

rpart.plot(tree_fit, type = 4, extra = 1, cex = 0.8)
# Make predictions on training and testing data
train_pred <- predict(tree_fit, newdata = train_data, type = "class")
test_pred <- predict(tree_fit, newdata = test_data, type = "class")
#Confusion matrix
# Convert train_pred to factor with levels c(10, -10)
train_pred_factor <- factor(train_pred, levels = c(10, -10))
# Convert train_data$Default.History to factor with levels c(10, -10)
y_train_factor <- factor(train_data$Default.History, levels = c(10, -10))
# Calculate confusion matrix for training data
train_confusion_decision_tree <- confusionMatrix(train_pred_factor, y_train_factor)
# Print confusion matrix
print(train_confusion_decision_tree)
# Convert test_pred to factor with levels c(10, -10)
test_pred_factor <- factor(test_pred, levels = c(10, -10))
# Convert test_data$Default.History to factor with levels c(10, -10)
y_test_factor <- factor(test_data$Default.History, levels = c(10,-10))
# Calculate confusion matrix for test data
test_confusion_decision_tree <- confusionMatrix(test_pred_factor, y_test_factor)

# Print confusion matrix
print(test_confusion_decision_tree)

# Print performance metrics
cat("Training Accuracy:", train_accuracy, "\n")
cat("Training F1 Score:", train_f1_score, "\n")

cat("Testing Accuracy:", test_accuracy, "\n")
cat("Testing F1 Score:", test_f1_score, "\n")

# Calculate quality metrics
train_auc <- roc(train_data$Default.History, as.numeric(train_pred))$auc
test_auc <- roc(test_data$Default.History, as.numeric(test_pred))$auc

```

```

# Print performance metrics
cat("Training AUC:", train_auc, "\n")
cat("Testing AUC:", test_auc, "\n")
# Plot ROC curve for testing data
test_roc <- roc(test_data$Default.History, as.numeric(test_pred))
plot(test_roc, print.thres = "best", main = "ROC Curve")

```

### **#Random forests-----**

```

# Load libraries
library(tidyverse)
library(randomForest)
library(caret)
library(pROC)
# Fit random forest
rf_fit <- randomForest(Default.History ~ ., data = train_data, ntree = 500, importance
= TRUE)
# Plot variable importance
varImpPlot(rf_fit)
# Make predictions on training and testing data
train_pred <- predict(rf_fit, newdata = train_data)
test_pred <- predict(rf_fit, newdata = test_data)
#Confusion matrix calculation
# Convert train_pred to factor with levels c(10, -10)
train_pred_factor <- factor(train_pred, levels = c(10, -10))
# Convert train_data$Default.History to factor with levels c(10, -10)
y_train_factor <- factor(train_data$Default.History, levels = c(10, -10))
# Calculate confusion matrix for training data
train_confusion_rf <- confusionMatrix(train_pred_factor, y_train_factor)
# Print confusion matrix for training data
print(train_confusion_rf)
# Convert test_pred to factor with levels c(10, -10)
test_pred_factor <- factor(test_pred, levels = c(10, -10))

```

```
# Convert test_data$Default.History to factor with levels c(10, -10)
y_test_factor <- factor(test_data$Default.History, levels = c(10, -10))
# Calculate confusion matrix for test data
test_confusion_rf <- confusionMatrix(test_pred_factor, y_test_factor)
# Print confusion matrix for test data
print(test_confusion_rf)
# Print AUC score
cat("Testing AUC:", test_auc, "\n")
# Plot ROC curve
plot(test_roc, print.thres = "best", main = "ROC Curve")
# Plot random forest
plot(rf_fit)
```



