

**ENHANCED SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE
(SMOTE) BASED MODEL FOR ENHANCING ACCURACY IN CREDIT
MODELLING**

DICKSON MURITHI

**A Thesis Submitted to Graduate School in Partial Fulfilment of the Requirements
for the Award of the Degree of Master of Science in Computer Science of Tharaka
University.**

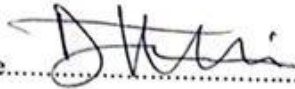
THARAKA UNIVERSITY

NOVEMBER 2024

DECLARATION AND RECOMMENDATION

Declaration

This thesis is my original research and has not been submitted for conferment of a degree in any other institution.

Signature.....

Date.....22/11/24

Dickson Murithi
SMT22/03135/21

Recommendations

This thesis has been submitted for examination with our approval as the University supervisors;

Signature.....

Date.....22/11/2024

Mr. Alexander Muriithi
Department of Computer Science
Tharaka University

Signature.....

Date.....22/11/24

Mr. Michael Mutisya
Department of Computer Science
Tharaka University



COPYRIGHT

©2024

All rights reserved. No single part of this thesis may be transmitted or reproduced in any form or any means, including mechanical or electronic methods, such as photocopying, information storage, recording, and retrieval systems, without prior written permission of the author or Tharaka University.

DEDICATION

Every work is a challenge and needs effort and guidance including by those very close to our hearts. I dedicate this work to my son Wayne, my mum Anita and my dad Julius whose love, prayers and encouragement made me able to get such success and honour.

ACKNOWLEDGEMENT

I am grateful and thankful to God for enabling me complete the work. I would also like to appreciate my ever-committed supervisors Mr. Alexander Muriithi & Mr. Michael Mutisya for their support, patience and knowledge. Their advice helped and inspired me throughout all my research. To my friend and mentor in Machine learning, Daniel Kilemi, thanks a lot. I also appreciate the contributions by other members of the department of computer science who ensured that we were enlightened on the requirements of the thesis and how to make effective contributions in research. I appreciate interactions and useful discussions that I had with the other graduate students at the department of computer science throughout our entire study.

Finally, this research was also made possible through the contributions of many other individuals all of whom I may not be able to name and I really acknowledge them for their input, guidance and help in any way. Special mentions to my friends Morris, Collins, Martin, Dickens and Noel.

ABSTRACT

Credit modelling especially in the financial sector faces significant challenges in deep learning applications. Accurate credit modelling is essential in enabling financial institutions assess credit risk and make viable lending decisions. However, complexities arise due to inaccuracies influenced by datasets, modelling algorithms, and sampling techniques. This research sought to evaluate and validate the effectiveness of the enhanced Synthetic Minority Oversampling Technique (SMOTE) based model in enhancing accuracy in credit Modelling. The enhanced SMOTE-based model integrates traditional and machine learning methods, including decision trees, logistic regression, Neural Networks, Random Forest, and Support Vector Machines. Using a diverse dataset, it incorporates borrower characteristics like age, income, and credit score, and loan details such as amount, interest rate, and term. The model focused on balancing data distribution, creating synthetic samples, addressing the challenge of overfitting and optimizing performance to surpass baseline models across metrics like accuracy, precision, recall. Findings revealed that there was limited adoption of advanced models amongst financial institutions in Meru County, due to their complexity and training demands. Further findings reveal that, applying enhancements to SMOTE based model improved class balance, accuracy and error reduction. Random Forests demonstrated marked improvements with enhanced model. Accuracy increased from 59.19% to 87.70%, and the Kappa statistic improved from 0.0055 to 0.7249, indicating better classification agreement. Error rates showed significant reductions, with the mean absolute error decreasing from 40.81% to 12.30%, and the root mean squared error dropping from 0.6388 to 0.3507. The enhancements in sensitivity (from 78.28% to 91.19%) and specificity (from 22.22% to 80.95%) further underscore the model's effectiveness in handling dataset imbalances with SMOTE. These results suggest that Random Forests, when combined with enhanced SMOTE-based models, can significantly improve the accuracy and precision of credit risk predictions. Adopting enhanced SMOTE-based models with Random Forests offers robust tools for credit risk management, advocating for increased quantitative model adoption and collaboration among financial institutions.

TABLE OF CONTENTS

DECLARATION AND RECOMMENDATIONError! Bookmark not defined.

COPYRIGHT **iii**

DEDICATION..... **iv**

ACKNOWLEDGEMENT..... **v**

ABSTRACT..... **vi**

TABLE OF CONTENTS **vii**

LIST OF FIGURES **xii**

LIST OF TABLES **xiii**

LIST OF ABBREVIATIONS **xiv**

CHAPTER ONE **1**

INTRODUCTION..... **1**

 1.0 Introduction 1

 1.1 Background of the Study..... 1

 1.2 Statement of the Problem 4

 1.3 Purpose of the Study 5

 1.4 Objectives of the Study 5

 1.5 Research Questions 5

 1.6 Significance of the Study 6

 1.7 Scope of the Study..... 6

 1.8 Limitations of the Study 7

 1.9 Expected Output..... 7

 1.10 Operational Definition of Terms 8

CHAPTER TWO	9
LITERATURE REVIEW	9
2.0 Introduction	9
2.1 Traditional Credit Risk Assessment.....	9
2.1.1 Linear Regression	9
2.1.2 Discriminant Analysis.....	9
2.1.3 Probit Analysis and Logistic Regression	10
2.1.4 Judgment-Based Models.....	10
2.2 Machine Learning Approaches in Credit Scoring.....	10
2.2.1 Decision Trees.....	11
2.2.2 Random Forests	11
2.2.3 Logistic Regression.....	12
2.2.4 Artificial Neural Network	12
2.2.5 Support Vector Machines.....	13
2.3 Related Research Work.....	13
2.3.1 Overview of Credit Modelling Techniques.....	13
2.3.2 Overview of SMOTE Models.....	19
2.3.3 Effectiveness of SMOTE Based Models in Enhancing Credit Modelling Accuracy	21
CHAPTER THREE	27
RESEARCH METHODOLOGY	27
3.0 Introduction	27
3.1 Location of the Study	27
3.2 Research Design.....	27
3.3 Population.....	29

3.4 Sampling.....	29
3.5 Sample Size.....	30
3.6 Data Collection.....	30
3.7 Conceptual Design	32
3.8 Proposed Model.....	33
3.9 Data Preprocessing.....	36
3.9.1 Data Cleaning.....	37
3.9.2 Data Reduction.....	38
3.9.3 Feature Engineering.....	39
3.9.4 Exploratory Data Analysis	40
3.9.5 Converting Categorical Variables	41
3.9.6 Standard Scaler	42
3.9.7 Handling Outliers.....	42
3.9.8 Modelling.....	45
3.9.9 Model Testing	47
3.10 Performance Metrics	50
3.10.1 Accuracy	50
3.10.2 Precision.....	50
3.10.3 Recall	50
3.10.4 Specificity	51
3.10.5 F1 Score	51
3.6 Ethical Considerations.....	51
CHAPTER FOUR.....	52
DATA ANALYSIS, FINDINGS AND RESULTS.....	52
4.0 Introduction	52

4.1 Demographic Presentation of Data	52
4.1.1 Distribution of Loan Amount.....	52
4.1.2 Rate of Default.....	53
4.1.3 Reason for Borrowing a Loan.....	54
4.1.4 Gender.....	55
4.1.5 Age Distribution of Customers	55
4.1.6 Collateral Offered to Secure Credit	56
4.2 Credit Models in Use by Financial Institutions.....	57
4.2.1 Methods Used for Accessing Credit Risk by Financial Institutions	57
4.2.2 Use of Quantitative Models in Modelling Credit Risk	60
4.3 Proposed Enhanced SMOTE-based Model.....	62
4.3.1 Data Preparation.....	62
4.3.2 SMOTE Implementation.....	62
4.3.3 Addressing overfitting.....	62
4.3.4 Model Selection and Training.....	63
4.4 Performance Metrics	63
4.5 The enhanced SMOTE-based model results	63
4.6 Effectiveness of Enhanced SMOTE-based Model in Enhancing Credit Modelling Accuracy.....	66
4.6.1 Comparison of Logistic Regression Algorithm Incorporating SMOTE	66
4.6.1.1 Comparison of Accuracy, Sensitivity & Specificity of Logistic Regression with Standard SMOTE Model and Enhanced SMOTE-based Model	68
4.6.2 Comparison of Decision Trees Algorithm Incorporating SMOTE	69
4.6.2.1 Comparison of Accuracy, Sensitivity & Specificity of Decision Trees with Standard SMOTE Model and Enhanced SMOTE-based Model	71
4.6.3 Comparison of Random Forests Algorithm Incorporating SMOTE.....	72

4.6.3.1 Comparison of Accuracy, Sensitivity & Specificity of Random Forests with Standard SMOTE Model and Enhanced SMOTE-based Model	74
4.6.3.2 Importance of Predictors Used in the Random Forests Model	75
4.6.4 Comparison of Support Vector Machines Algorithm Incorporating SMOTE .	77
4.6.4.1 Comparison of Accuracy, Sensitivity & Specificity of SVM with Standard SMOTE Model and Enhanced SMOTE-based Model.....	78
4.6.5 Comparison of Artificial Neural Networks Algorithm Incorporating SMOTE	80
4.6.5.1 Comparison of Accuracy, Sensitivity & Specificity of ANN with Standard SMOTE Model and Enhanced SMOTE-based Model.....	81
CHAPTER FIVE	83
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	83
5.0 Introduction	83
5.1 Summary of Key Findings	83
5.2 Conclusions	86
5.3 Recommendations	87
REFERENCES.....	89
APPENDICES	96
Appendix 1: Sample Questionnaire.....	96
Appendix 2: A sample of Importing Libraries	100
Appendix 3: A sample of Creating Random Samples.....	101
Appendix 4: Metrics Calculation	102
Appendix 5: Tharaka University Introductory Letter	104
Appendix 6: Ethics Review Letter	105
Appendix 7: NACOSTI License.....	106
Appendix 8: Meru County Map.....	107

LIST OF FIGURES

Figure 1.1: Synthetic Sample generation using SMOTE (Ahmad,et al., 2023).....	3
Figure 2.1: Artificial Neural Networks	13
Figure 3.1: Research Design	29
Figure 3.2: Stratified Sampling.....	30
Figure 3.3: CRISP-DM Methodology.....	32
Figure 3.4: Proposed Model.....	33
Figure 3.5: Importing Python Libraries	36
Figure 3.6: Train Data	36
Figure 3.7: Removing Missing Values	37
Figure 3.8: Preprocessed Data	38
Figure 3.9: Creating Target Variable	39
Figure 3.10: Converting Categorical Variables.....	41
Figure 3.11: Scaling data.....	42
Figure 3.12: Normalized Income Total	43
Figure 3.13: Normalized Amount	44
Figure 3.14: Variable Declaration	45
Figure 3.15: Splitting Data.....	45
Figure 3.16: SMOTE algorithm.....	46
Figure 3.17: Training a Decision Tree Classifier	46
Figure 3.18: Figure fitting.....	47
Figure 3.19: Test Data Preprocessing	48
Figure 3.20: Handling Outliers in Test Data	49
Figure 3.21: Loan Default Prediction	49
Figure 4.1: Distribution of Loan Amount	53
Figure 4.2: Distribution of customers Age.....	56
Figure 4.3: ROC Curves Comparing Different Models.....	65

LIST OF TABLES

Table 1: Rate of Loan Default.....	53
Table 2: Reason for Borrowing Loan.....	54
Table 3: Gender of Customers	55
Table 4: Collateral.....	57
Table 5: Methods of credit modelling.....	58
Table 6: Methods of Accessing Credit Risk.....	60
Table 7: Comparison of Accuracy, Sensitivity & Specificity of standard SMOTE with Enhanced SMOTE-based model. (Before enhancements and after Enhancements)	64
Table 8: Comparison of Logistic Regression Algorithm Metrics with standard SMOTE model and enhanced SMOTE-based model.....	67
Table 9: Comparison of Accuracy, Sensitivity & Specificity of SMOTE with Logistic Regression.....	68
Table 10: Comparison of Decision Trees Model Metrics with standard SMOTE model and enhanced SMOTE-based model.....	70
Table 11: Comparison of Accuracy, Sensitivity & Specificity of SMOTE with Decision Trees.....	71
Table 12: Comparison of Random Forests Model Metrics with standard SMOTE and enhanced SMOTE-based model	73
Table 14: Importance of Predictors Used in the Random Forests Model	76
Table 15: Comparison of SVM Model Metrics with standard SMOTE model and enhanced SMOTE-based model	77
Table 16: Comparison of Accuracy, Sensitivity & Specificity of SVM with standard SMOTE and enhanced SMOTE-based model	79
Table 17: Comparison of ANN Model Metrics with standard SMOTE model and enhanced SMOTE-based model	80
Table 18: Comparison of Accuracy, Sensitivity & Specificity of ANN with Standard SMOTE Model and Enhanced SMOTE-based Model.....	82

LIST OF ABBREVIATIONS

ADASYN	Adaptive Synthetic Sampling
ANN	Artificial Neural Networks
AUC	Area Under the Curve
CART	Classification and Regression Trees
CDSMOTE	Categorical Data SMOTE
CNN	Convolutional Neural Network
DA	Data Augmentation
FICO	Fair Isaac Corporation
LSTM	Long Short-Term Memory (a type of recurrent neural network)
NACOSTI	National Commission for Science, Technology and Innovation
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
WSMOTE	Weighted Synthetic Minority Over-sampling Technique

CHAPTER ONE

INTRODUCTION

1.0 Introduction

The chapter introduces a background for the study, purpose of the study and research questions that guide this study. It also demonstrates the significance of the research study its scope and limitations. It outlines the expected output of this study and operational definition of terms used.

1.1 Background of the Study

Credit modelling involves use of machine learning and statistical models to transform client data into actionable insights capable of informing credit decisions and action for any financial institution (Anderson, 2007). These models categorize applicants as good or bad debtors based on characteristics like income, age, and marital status. Credit modelling serves to lower the cost of credit by mitigating default risks through the evaluation of a customer's creditworthiness, and in certain cases, it aids in fraud detection. Moreover, the model has the capability to monitor ongoing loan accounts, enabling the prioritization of repayment collection. Presently, nearly all credit institutions engage in some type of credit modelling and assessment before extending lines of credit or loans to both individuals and corporations.

Despite Synthetic Minority Oversampling Technique (SMOTE) being powerful, it has limitations. One drawback is that it does not consider the underlying distribution in the minority class; potentially leading to the generation of noisy or unrealistic synthetic instances (Guidolin & Pedio, 2021). The problem of inaccuracies caused by imbalanced data (Hamal & Senvar, 2021), noise and uncertainty in credit Modelling has been widely recognized in literature. Resampling, in addition to cost-sensitive learning, as well as ensemble methods such as random are popular techniques used in machine learning to address these challenges. The methods have advantages and weaknesses, and the choice of the appropriate method depends on the problem under consideration and the available data.

Resampling is a strategy that involves modifying the training dataset to create a more balanced dataset (Jagelid & Movin, 2021). This can be done by either under-sampling the class with majority observations or oversampling the class with minority observations. The competitive advantage of resampling is that it is ideal for application in real scenario since it enhances the performance of a model in the presence of class imbalance (Ahmad, Kasasbeh, Al-dabaybah, & Faisal, 2023), it does not require additional data to be collected and it can be relatively simple to implement. The disadvantage of resampling is that it could lead to over-fitting, as the model may learn to memorize the training dataset rather than generalizing to a new dataset.

On the other hand, Cost-sensitive learning incorporates assigning different costs to different types of errors, depending on their severity. The cost-sensitive algorithms look to minimize the total cost-of-the-errors rather than the traditional approach of minimizing the classification error rate (Ling & Sheng, 2015). The drawback of cost-sensitive type of learning is that it requires domain expertise to assign appropriate costs to different types and forms of errors.

Ensemble methods combine and incorporate multiple models to enhance performance. The models can be trained on various subsets of the training dataset, using various algorithms. Ensemble techniques of creating models can reduce over-fitting and enhance generalization, as the diversity of the models reduces the risk of all models making the same errors (Abedin, Guotai, & Hajek, 2022). However, the disadvantage of ensemble methods is that they may be computationally expensive and difficult to interpret.

Based on the competitive advantage of resampling technique over other techniques, this study proposes the use of SMOTE in Modelling. SMOTE was developed by Chawla et al. in 2002 as an oversampling technique that creates synthetic samples for the minority class by interpolating between existing instances of the minority class (Guo, et al., 2017).

SMOTE has gained popularity in different domains, among them credit modelling, fraud detection, medical diagnosis, and image classification, where class imbalance is a common challenge (Johnson & Khoshgoftaar, 2019). It has been shown to effectively

improve classification accuracy, reduce bias, and enhance the performance of machine learning models.

The SMOTE method works by identifying the minority class instance together with its nearest neighbours, and then creating new synthetic instances on the line segments joining the instance to its other neighbours. The synthetic instances generated by SMOTE help to balance the distribution within the classes, allowing the machine learning algorithm to learn from a representative dataset. By increasing the minority class instances, SMOTE helps to alleviate the bias towards the majority classes and increases the overall performance of the classifier model in accurately predicting both classes (Ahmad, et al.,2023) as illustrated in figure 1.1.

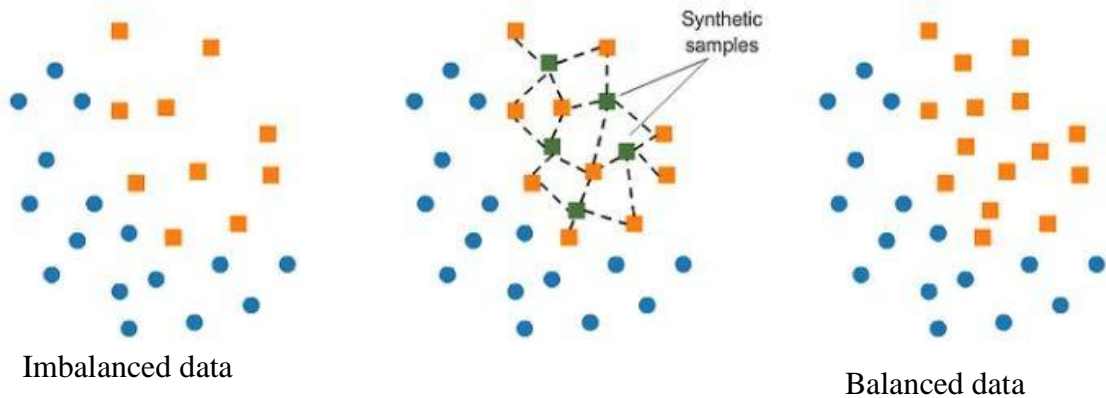


Figure 1.1: Synthetic Sample Generation using SMOTE (Ahmad,et al., 2023)

Despite SMOTE being powerful, it has its limitations. One drawback is that it does not consider the underlying distribution in the minority class, potentially leading to the generation of noisy or unrealistic synthetic instances. Additionally, SMOTE may not perform well when there are overlapping or inseparable clusters within the minority class. To address the limitations, researchers have put forth various extensions and modifications to SMOTE, such as Adaptive Synthetic Sampling (ADASYN), Borderline SMOTE, and SMOTE Boost, to further enhance its performance and adaptability to different datasets (Yong, et al., 2022).

In the context of credit modelling, SMOTE has been employed to handle imbalanced datasets and improve the accuracy of credit risk prediction (Elyan, Moreno-Garcia, & Jayne, 2020). By generating synthetic minority class instances, SMOTE helps to ensure that credit models are better equipped to identify and assess high-risk borrowers accurately.

1.2 Statement of the Problem

The accuracy of credit risk assessment models confronts a persistent challenge, exacerbated by imbalanced datasets within the realm of big data. Traditional approaches such as the Fair Isaac Corporation (FICO) scores and statistical methods, often struggle to effectively capture the minority class, resulting in biased and unreliable predictions. This limitation in credit Modelling poses significant threats to financial institutions, leading to elevated default rates and potential financial losses. Although modern techniques, including cost-sensitive learning, ensemble methods, and resampling, have been employed to address these challenges, each approach has its own disadvantages. Cost-sensitive learning, while aiming to minimize error costs, necessitates domain expertise for assigning appropriate costs to different error types. Ensemble methods enhance performance but are computationally expensive and challenging to interpret. Resampling, a method of modifying training data to create a more balanced dataset, is relatively simple to implement, making it ideal for real-world applications. However, it can lead to overfitting. SMOTE addresses imbalanced datasets by generating synthetic samples but falls short in capturing complex relationships and overlapping regions between classes. This results in the introduction of noise and unrealistic instances that disrupt the balanced data distribution, hindering the optimization of synthetic samples and contributing to overfitting issue. To tackle the challenge of imbalanced data and overfitting this research aimed to develop an enhanced SMOTE-based model. The model explored advanced variations of SMOTE and integrated additional techniques which include data pre-processing methods, feature selection, ensemble learning, and model comparisons. SMOTE and overfitting algorithms were applied to standard SMOTE model as further enhancements. By incorporating these enhancements into the resampling process, this development created a more robust and accurate credit Modelling framework. This

framework was designed to effectively capture the minority class, thus significantly improving overall predictive performance. The successful implementation and development of this model is anticipated to empower financial institutions to make well-informed lending decisions, reduce default rates, and enhance the stability and reliability of credit risk assessments.

1.3 Purpose of the Study

The purpose of the study was to develop an enhanced SMOTE-based model that effectively addresses imbalanced data challenges and overfitting. By leveraging SMOTE and integrating supplementary methodologies, this research aims to improve the prediction accuracy of all credit risk models. Ultimately, the study seeks to empower financial institutions to make more viable lending decisions, thereby reducing default rates and enhancing the reliability of credit risk assessments.

1.4 Objectives of the Study

- i. To investigate through a survey the existing credit modelling techniques.
- ii. To develop an enhanced SMOTE-based model that enhances accuracy in credit modelling.
- iii. To assess the effectiveness of the developed model in enhancing credit modelling accuracy using lab data.

1.5 Research Questions

- i. What existing credit modeling techniques are currently utilized and how do they perform?
- ii. What enhancements can be incorporated to standard SMOTE model to enhance its accuracy in credit modelling?
- iii. What is the effectiveness of the enhanced SMOTE-based model in enhancing credit modeling accuracy using lab data?

1.6 Significance of the Study

The study exhibits great significance in the following aspects.

- **Enhancement of Credit Models:** Aims to increase the accuracy- predictive quality of credit models by effectively addressing challenges posed by imbalanced datasets, a common issue in credit-risk assessment.
- **Contribution to Research:** Provides valuable insights and potential solutions for ongoing research in enhancing credit modeling methodologies.
- **Relevance to Financial Industry:** Enables financial institutions e.g. banks and credit consumers to make more informed and better decisions regarding lending and borrowing risks.
- **Better Risk Assessments:** Improved accuracy in credit modeling can lead to reduced likelihood of defaults and financial losses for lending institutions.
- **Clarity for Borrowers:** Offers borrowers a clearer understanding of their creditworthiness

Overall, the significance of this study is majorly in its potential to contribute to the advancement of credit risk assessment practices, ultimately fostering more secure and efficient lending practices and benefiting both financial institutions and credit consumers alike.

1.7 Scope of the Study

The scope of the study was limited and constrained to the use of the SMOTE technique in credit modelling. The research was based on Kenyan dataset from top performing Meru County lending institutions. The datasets are proprietary in nature and requires strict adherence to confidentiality guidelines. Data anonymization was done to protect the owners from being exposed. The research evaluated the effectiveness of SMOTE in enhancing the accuracy and predictive performance of credit models compared to other techniques and traditional methods such as logistical regression models. The study used python as our primary programming language and jupyter lab as our coding tool. Python libraries including Scikit-learn, Pandas, and Numpy for data preprocessing, feature

extraction and machine learning. SMOTE used imbalanced learn libraries to perform predictions.

1.8 Limitations of the Study

The study came face to face with several notable limitations. First, the accuracy of the credit model is contingent on the availability and the quality of the credit datasets, potentially affecting the model's performance if the data is biased or contains errors. The limitation was overcome by applying data preprocessing methods such as data-cleaning, data reduction and feature engineering. Secondly, while the study evaluated various machine learning algorithms and resampling techniques, all Modelling approaches were not to be considered. Additionally, the computational resources available limited the number of synthetic samples generated using SMOTE, impacting the model's performance. Lastly, as a retrospective analysis, the study could not predict future credit trends or assess the model's real-world effectiveness. Future studies should evaluate model effectiveness in real-world scenarios through longitudinal studies, real-world data validation, and case analyses. Collecting user feedback and conducting comparative analyses will enhance understanding. Additionally, scenario testing and assessing regulatory compliance can shed light on the model's adaptability and impact on financial decision making.

1.9 Expected Output

- i. Development of an enhanced SMOTE-based model for enhancing accuracy in credit modeling. This output involves creating a robust model that addresses data imbalance, improving predictive accuracy.
- ii. Thesis. A comprehensive document detailing research findings, methodologies, and contributions to the field.
- iii. Publication in a refereed journal. Peer-reviewed article sharing research results, validating the study's contributions to academia.
- iv. Presentation of research findings at learned conferences. Opportunities to share insights and learning and engage with field experts, fostering collaboration and discussion.

1.10 Operational Definition of Terms

- Logistic Regression:** A classification technique and algorithm that predicts binary outcomes using a logistic-function to map predictions to a range between 0 and 1. It's used to determine which of two classes an instance belongs to based on predictor variables.
- Machine Learning:** A field of AI that enables systems and machines to learn from data and make predictions based using data.
- Neural Networks:** Models mirroring the brain's structure that learns to recognize patterns through interconnected layers of nodes.
- Overfitting:** A problem where a model learns the training dataset too well, including noise and errors, leading to poor generalization to new data.
- Random Forest:** A model that builds multiple and many decision trees and incorporates their combined outputs to increase prediction accuracy and minimize overfitting.
- Resampling:** Techniques used to adjust dataset size and composition, often to evaluate model performance or address class imbalance.
- SMOTE:** It is method for addressing imbalance classes by creating synthetic-samples for the minority class.
- Support Vector Machine:** A classification and regression algorithm that seeks to optimize the decision boundary by maximizing the margin between classes.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

This chapter reviews traditional statistical methods for credit-risk evaluation, such as logistic regression and decision trees, as well as advanced machine learning algorithms, like neural networks and random forests. It also explores the use of SMOTE to address and tackle class imbalance in credit datasets and discusses the implications of these methods for the specific research problem.

2.1 Traditional Credit Risk Assessment

Traditional credit scoring models typically employ historical data, including bank transaction data (past credit utilization, repayment behavior), credit bureau reports (payment history, credit inquiries), and commercial data (financial ratios, business age) (World Bank, 2020). Linear regression, discriminate analysis, judgment-based models, probit analysis and logistic regression are commonly used traditional credit models.

2.1.1 Linear Regression

Linear regression is a classical statistical method used to model a linear relationship between a dependent or target variable (y) and one or more explanatory or independent variables (x_1, x_2, \dots, x_p). A linear regression model is expressed in the form: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$ where y is the response variable, x_1, x_2, \dots, x_p are the explanatory variables, $\beta_0, \beta_1, \dots, \beta_p$ are regression coefficients and ε (random) is the error term or random effect. Before constructing a linear regression model, it's essential to assess the correlation between some of the explanatory variables of interest to ensure a meaningful relationship exists.

2.1.2 Discriminant Analysis

Discriminant Analysis is a statistical technique developed by Sir Ronald Fisher in the year 1936 to classify observations into predefined groups. A common application in credit scoring is to classify borrowers as default or non-default. Linear Discriminant Analysis (LDA) was one of the earliest statistical methods used to predict corporate

bankruptcy. It employs financial ratios and other relevant indicators to distinguish between firms that are likely to default and those that are not.

2.1.3 Probit Analysis and Logistic Regression

The probit and logit models are statistical models used to model the probability of a binary outcome. In the Probit Model, the inverse standard normal distribution of the probability is modeled as a linear combination of predictor variables. On the other hand in the logit model, The log-odds of the probability is modeled as a linear combination of predictor variables. The logit model can be expressed as: $\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n$, where p is the probability of the event occurrence, the β_i are the model coefficients and x_i are the covariates.

2.1.4 Judgment-Based Models

According to the authors Bana e Costa, Barroso, and Soares (2002), the analytic hierarchy process (AHP) is a well-structured approach for analyzing and organizing complex decisions. This method involves breaking down a complex and big decision problem into a hierarchy of simpler sub-problems, each of which can be evaluated individually. A key feature of AHP is its incorporation of human judgment, alongside quantitative data, to inform the evaluation process.

2.2 Machine Learning Approaches in Credit Scoring

Machine learning involves developing algorithms that learn from data and make decisions or predictions. There are three types of machine learning, one, supervised learning, two, unsupervised learning and three reinforcement learning. In supervised learning the algorithm is fitted and trained on a labeled dataset, where each data point is associated with a correct output. The goal is to learn a mapping unique function that can accurately and simply predict the output for new, unseen data. In contrast, unsupervised learning algorithms are trained and fitted on unlabeled data. The goal is to discover hidden patterns or structures within the data. Common tasks include clustering, where data points with similar characteristics are grouped together, and association rule mining, where relationships between different variables are identified. Finally, in reinforcement learning, an agent is trained through interactions with an environment of interest. At each

interaction, the agent is rewarded based on actions either by penalizing a wrong response and giving more weight to a correct action. By learning from trial and error, the agent can develop optimal strategies. Decision trees, logistic regression models, support vector machines (SVM) and other associated models, Artificial neural networks (ANN) and random forests (RF) were the machine learning algorithms that were reviewed and used by researcher in this research to assess their ability to enhance accuracy in credit modelling when combined with SMOTE.

2.2.1 Decision Trees

Decision Trees are supervised learning models that are suitable for classification and regression use cases. They represent a decision-making process in a tree-like structure, in which internal nodes correspond to some attribute, while each branch represents a value of that attribute, and each leaf node of the tree represents a class label or a predicted value. Categorical decision trees are used when the target or response variable is categorical, such as "yes" or "no," or "spam" or "not spam." The tree splits the data based on the different values of the attributes to create homogeneous or similar subsets, where most instances belong to the same class. In contrast, continuous variable decision trees are used when the response variable is continuous, such as predicting real estate prices or stock prices. The tree splits the data based on attribute values to create subsets where the target variable values are as similar as possible. By following the decision rules from the root node to a leaf node, we can make predictions and correct judgments for new, unseen before data.

2.2.2 Random Forests

Random Forests are powerful machine learning algorithms that leverage the ensemble learning technique of bagging. It constructs a multitude of decision trees, each of the trees trained on a random subset of the original dataset. By averaging the predictions of these individual trees, Random Forest achieves superior accuracy and robustness compared to single decision trees. This versatility makes it applicable to both classification and regression tasks, making it a valuable tool for various data analysis and predictive modeling applications.

2.2.3 Logistic Regression

Logistic regression (LR) is a classification statistical model used to model data and understand the relationship between a binary outcome variable and one or more explanatory variables. When using binary logistic regression, a number of key assumptions must be met.

First, the dependent or response variable must be binary, such as a "yes" or "no" outcome. Second, the data should not contain significant outliers. This can be checked by standardizing continuous predictors to identify and address any extreme values. Lastly, the independent variables should not exhibit strong correlations with one another, a condition known as multicollinearity. To assess this, a correlation matrix of the predictors is created to ensure the variables are largely independent.

2.2.4 Artificial Neural Network

Artificial Intelligence encompasses neural networks, a class of machine learning models inspired by the behavior of biological neurons. A neural network consists of nodes that process incoming data and pass the results to the subsequent nodes. The output of each node, referred to as its activation or node value, is influenced by diverse weights assigned to the nodes. These weights, which can be adjusted, determine the strength of an input's impact on the output and enable the network to learn. Activation functions, such as linear, sigmoid, ramp, hyperbolic, or Gaussian, are used in performing the network's calculations. The Multilayer Perceptron Model, capable of recognizing non-linear patterns, is particularly effective in applications such as fraud detection. Figure 2.1 illustrates how ANN Multilayer Perceptron works.

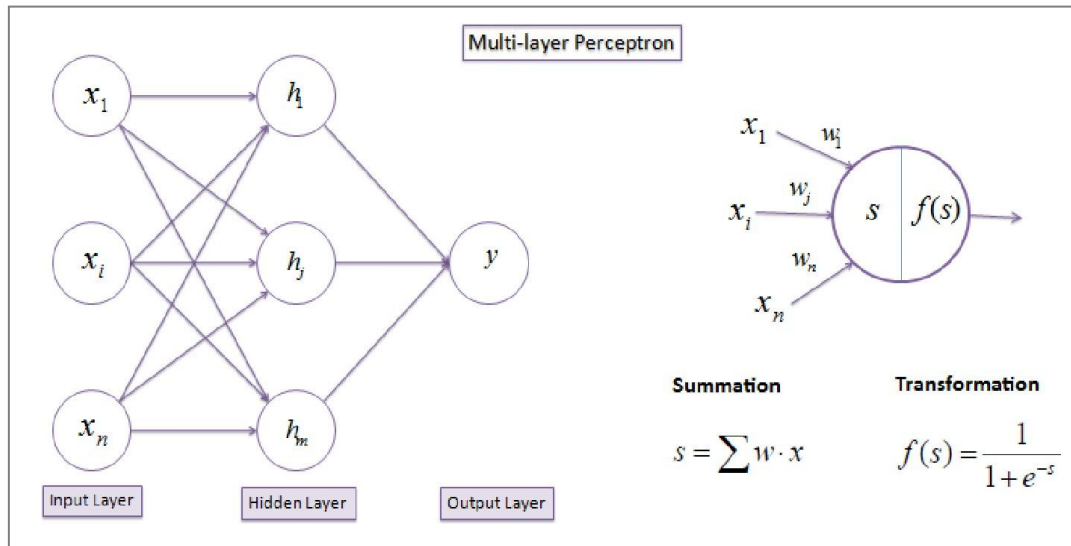


Figure 2.1: Artificial Neural Networks

2.2.5 Support Vector Machines

Support Vector Machine (SVM) is a powerful supervised learning algorithm widely used for classification tasks, such as predicting credit risk. It works by finding the optimal hyperplane that divides the data points of different classes, such as bad or good credit risks. SVM finds a local maximum between data points near to each other (support vectors) of each class. In credit modelling assessment, features like income, credit history, and debt levels are used as input for predicting the probability of default. By training on historical data, SVM can classify new applicants into risk categories, helping lenders make informed decisions. Its capabilities in handling non-linear relationships make it particularly effective in this domain.

2.3 Related Research Work

2.3.1 Overview of Credit Modelling Techniques

Credit is a fundamental aspect of commercial banks, microfinance institutions, SACCOS and other financial institutions that offer myriad services like giving business loans, taking deposits, and offering basic investment products. However, granting loans and credit to customers inherently involves taking risks, as these activities play a key role in the bank's economic growth. Striking the right balance is crucial, as both overly cautious and excessively risk-taking approaches can jeopardize a bank's survival (Mandala, Nawangpalupi, & Praktikto, 2012). Hence, effective risk management for banks

necessitates the identification, understanding, measurement, and implementation of appropriate strategies to issue out credit to their customers, without being too cautious or without over issuance of credit that may be detrimental to the institution (Bekhet & Eletter, 2014).

Accordingly, credit risk is a major risk that each financial institution has to deal with. It refers to the likelihood of customers failing to make payments on time or being unable to repay loans (Cisko & Klieštik, 2013). To manage this risk, banks categorize customers into good and bad customer groups based on their probability of loan repayment (Akkoc, 2012).

The evaluation of credit risk holds utmost importance for banks, as they must ensure borrowers' ability to meet instalment payments before granting loans (Mandala, Nawangpalupi, & Praktikto, 2012). Adhering to Basel 2 guidelines, every lending institution must create an internal credit scoring system to assess the risks associated with borrowers. This has resulted in an increased need for precise scoring systems capable of Modelling risks comprehensively, leading institutions to formulate tailored models upon request from banks. These credit-scoring techniques function as decision tools or decision steps or algorithms, catering to different customers (Heiat, 2012).

Numerous models modelling credit-risk have been crafted through trial and error, lacking a robust theoretical foundation (Wang, Ma, & Yang, 2014). In addition, the models often exhibit a static nature and face challenges in maintaining effectiveness during economic crises.

Historically, banks have depended on static Modelling frameworks for evaluating customer credit risks. Yet, the limited adaptability of these models to evolving economic conditions renders them inefficient, especially when confronted with concept drifts. In instances where previously reliable customers might default, traditional static models, although reasonably effective in stable periods, prove inadequate in navigating the complexities introduced by economic and political fluctuations. This research therefore reviewed the success and challenges of some of the traditional static models used in

credit modelling such as linear regression, Discriminant analysis, judgement-based models and probit analysis.

Linear regression is widely used in consumer lending to predict default probabilities based on borrower characteristics such as income, credit score, and existing debts. For example, Huang and Chen (2020) applied linear regression to a dataset of 10,000 loan applications from a regional bank. The linear regression model achieved an accuracy rate of approximately 75% in predicting defaults. It effectively illustrated the relationship between income and credit scores versus default likelihood. However, the model's reliance on the assumption of linearity limits its effectiveness, especially in cases where borrower behavior is influenced by non-linear factors. Additionally, it is sensitive to outliers, which can skew results (Crook et al., 2007). To enhance predictive accuracy, it is recommended to incorporate more complex models, such as polynomial regression models or machine learning techniques, which can capture non-linear relationships better.

Discriminant analysis, notably the Altman Z-score, has been employed to assess bankruptcy risk in corporate finance. Altman (1968) used a dataset of 66 manufacturing firms, categorizing them based on financial ratios. The Z-score model demonstrated an impressive accuracy rate of 90% in distinguishing between distressed and non-distressed firms, proving effective across various industries. Discriminant analysis uses the underlying assumption that the data follows a multivariate normal distribution (MVN) and that the covariance matrices within each group are equal. This assumption might lead to misclassifications, particularly in cases of non-traditional borrowers (Ohlson, 1980). To improve robustness, it is advisable to incorporate non-parametric methods or ensemble techniques that do not rely on strict distributional assumptions, thereby enhancing model flexibility.

Probit analysis is frequently used in consumer finance and microfinance sectors to model default risk. Long and Freese (2006) applied this approach to a dataset of 5,000 microfinance borrowers, focusing on various predictor variables. Probit models achieved accuracy scores between 75% and 85%, effectively capturing the complexities of borrower behaviors in diverse populations. The complexity of probit models requires a

solid understanding of statistical theory, which may limit their accessibility for practitioners without strong quantitative backgrounds. Moreover, poor variable selection can lead to model overfitting (Bauer & Tharakan, 2006). It is recommended to conduct thorough exploratory data analysis (EDA) and feature selection before modeling. Using automated machine learning tools can also help ferret the most significant predictors while minimizing or reducing the risk of overfitting.

Judgment models integrate expert opinions alongside quantitative assessments in commercial lending. Gul and Kahn (2018) analyzed a dataset of 1,200 loan applications from small businesses, focusing on both qualitative and quantitative metrics. The judgment models were reported to have an accuracy rate of 75% to 80% in assessing credit risk, benefiting from the holistic view provided by expert evaluations. This subjective nature of judgment models may cause variability in assessments, as different analysts may interpret data differently. This inconsistency can reduce the reliability of risk evaluations (McKinsey & Company, 2015). To mitigate subjectivity, it is suggested to establish standardized frameworks for evaluating qualitative factors. Additionally, combining judgment models with statistical techniques can create a more robust risk assessment framework.

Consequently, multiple researchers have addressed the challenge of customer credit risk assessment through varied approaches and techniques, each endeavouring to provide a more precise model than its counterparts. Advancements in computer technology have significantly enhanced the ease of data collection and manipulation, increasing the demand for effective data analysis techniques and classification. (Zanin, 2016). Machine learning- learning outcomes through data, and Data mining stand out as popular techniques in this domain. Data mining involves exploring data to identify concealed patterns and even relationships (Sumathi & Sivanandam, 2006).

Hence, researchers have persistently delved into the application of a range of machine learning methods for credit assessment. Such methods encompass decision trees (DT), Support vector machines (SVM), artificial neural networks (ANN), random forests (RF) and integration algorithms (Correa, 2016) observed that binary logistic regression models

were not as effective as the aforementioned techniques when dealing with complex nonlinear relationships among various characteristic variables. While the Logistic Regression model may not achieve the prediction accuracy of some machine learning models, it offers distinct advantages in terms of variable interpretability and stability. As a result, certain researchers have enhanced the Logistic Regression model and employed it in predicting borrowers' default behaviour.

In (Masmoudi, Masmoudi, Abid, & Masmoudi, 2019a) they employed a discrete Bayesian network that incorporates potential vectors to forecast user payment defaults. The study by (Caruso, Gattone, Fortuna, & Di Battista, 2021) highlighted the relationship (correlation) between quantitative and the qualitative attributes of applicants, proposing a hybrid data-based segmentation analysis technique for credit risk assessment. On the other hand, (Li W. , Ding, Chen, Wang, & Yang, 2019a) introduced the notion of migration-learning for quick and automatic credit assessment. They imported data from traditional business to new business and crafted predictive models for credit evaluation.

A study by (Chopra & Bhilare, 2018) performed an empirical examination on a publicly accessible bank loan dataset, centring on banking loan default by employing decision trees as the foundational learner. The study compared the results of various ensemble tree learning techniques, such as boosting, bagging, and random forests. The findings revealed that the gradient boosting model outperformed the basic decision tree learner, demonstrating the superiority of ensemble methods over individual models. The study emphasized the importance of adopting modern techniques to improve accuracy and develop more effective loan recovery strategies.

In (Maha, Tony, & Niall, 2020) they employed cluster analysis on credit card account behaviour to assist in credit risk assessment. The behaviour of the accounts was parametrically modelled, and a recently advocated dissimilarity measure for statistical model parameters was applied in behavioural cluster analysis. This measure explicitly incorporated the uncertainty associated with parameter estimates derived from statistical models. The study led to the identification of noteworthy clusters within real credit card

behaviour data, enhancing the prediction as well as forecasting of account defaults based off on the clustering analysis outcomes.

The researchers (Shashi, Handa, & Singh, 2017) introduced a feature selection-based hybrid-bagging algorithm (FS-HB) to undertake credit risk assessment. The hybrid FS-HB algorithm outperformed standalone classifiers and reduced the type II error significantly. It showcased enhanced performance on datasets characterized by qualitative attributes, fewer features, and also tree-based unstable base classifiers. The algorithm's effectiveness was credited to its utilization of crucial features and the ensemble methodology, outperforming standalone classifiers in the balance between bias and variance.

In their study, (Kulkarni & Dhage, 2019) developed an innovative credit scoring system that merged Legacy credit scores, derived from financial history, with Emotional/Social credit scores, gathered from social media and web interactions. The proposed system, named "Information Trustworthiness," aimed to enhance data precision by cross-referencing with reliable sources. By incorporating an individual's personality traits, the Advanced Credit Score demonstrated superior accuracy in predicting financial behaviour when compared to the traditional Legacy credit score. Notably, the effectiveness of the Advanced Credit Score depended on selecting appropriate fractions from the Legacy credit score as well as Emotional/Social credit score. The introduced credit scoring system effectively differentiated individuals with a track record of defaults from those who had never utilized financial services like credit cards, a distinction challenging to achieve using conventional financial credit scoring methods.

For Chinese financial institutions dealing with competition from Internet financial businesses, (Li, Ding, & Chen, 2019) addressed this challenge by proposing an automatic credit-risk evaluation systems based off on machine learning. To overcome data scarcity for new businesses, they innovatively applied transfer learning, using data from traditional bank businesses to train defaulters risk prediction models on small data sets. The results demonstrated the commercial value of transfer learning techniques in the

financial-risk field, enabling effective risk management for emerging ventures amidst competitive pressures.

The study by (Moula, Guotai, & Abedin, 2017) focused on credit default prediction (CDP) Modelling, an important concern for financial institutions. Earlier, research indicated varying classifier performances in CDP analysis based on different databases and circumstances. However, the evaluation exercise using multiple performance criteria remained understudied, leading to inconsistent conclusions. The researchers addressed this problem by applying a support vector machines (SVM)-based CDP algorithm, incorporating a set of representative performance criteria, including novel measures. They compared its performance with statistical and intelligent approaches across 6 credit prediction databases. The experimental results demonstrated the SVM model's superiority, particularly over Classification and Regression Trees (CART) with Data Augmentation (DA), showcasing its robustness. Consequently, the study emphasized the importance of the evaluation metric in determining the classifier's supremacy.

2.3.2 Overview of SMOTE Models

The problem outlined in section 1.2 points to a necessary need to consider techniques of using imbalanced data to achieve reliable results. Often, the nature of any data used in a machine learning problem is rife with missing observations, imbalanced outcome classes, dimensionality problems and such other issues beyond the scope of this study. There have been successful attempts in literature to model data possessing imbalanced classes (Akkoc, 2012).

Class imbalance arises in the presence of a notable discrepancy in the number of instances belonging to two classes of the target variable. In this situation, one class is distinguished by a large number of instances, whereas the other class is represented by a comparatively limited number of instances (Japkowicz & Stephen, 2002). This issue has gained prominence across diverse fields that employ predictive models, such as Educational Data Mining (EDM). EDM utilizes data mining and machine learning techniques to tackle challenges within the educational domain (Kovács, 2019). In fact, other than education, class imbalances occur often in other real-world scenarios that

concern important occurrences of the population, objects, or area under study (Johnson & Khoshgoftaar, 2019). In the credit world for example, the number of defaulters may only constitute 10% of the customers issued credit. A reliable predictive algorithm in this case must accurately predict even the minority class in order to apply effective control measures.

However, several techniques in imbalanced classes learning have been created to address the class imbalance challenge through two main approaches: (1) alleviating bias that machine learning algorithms may show towards the majority observations class in a dataset, or (2) refining the algorithms to be more responsive to the minority class (Johnson & Khoshgoftaar, 2019). The methods encompass data-level approaches, emphasizing data resampling with various resampling techniques, and other algorithm-level approaches that prioritize adapting classification algorithms for improved handling of imbalanced datasets, and hybrid approaches that synergize the strengths of both data-level and by extension algorithm-level techniques to effectively address class imbalance (Kaur & Gosain, 2018). The study focuses on the first approach on data level imbalance learning.

Under the data level imbalance learning approach, there is a resampling of the target class to counter class imbalance. The resampling associated with synthetic minority oversampling (SMOTE) therefore oversamples the minority class in the target variable to match the majority class (Ma & He, 2013). The approach has been widely employed to counter class imbalance situations in practice, as earlier put.

In a comprehensive review conducted by (Guo, et al., 2017), out of 527 articles, it was noted that 156 (approximately 29.6%) employed resampling techniques to tackle issues of data imbalance. These articles covered a diverse array of disciplines, spanning at least 12 different fields. Additionally, (Kovács, 2019) performed an empirical study that entailed the comparison and evaluation of 85 over-sampling techniques in 104 datasets. Both investigations suggest that the efficacy of resampling techniques in addressing class imbalance problems have been thoroughly explored in contemporary literature. He stressed that the strength of each resampling technique will be dependent in not only on

its operating principle—how it balances the classes of the target variable—but also on the imbalance ratio between those classes. For example, when handling moderately imbalanced data, both oversampling and undersampling demonstrate similar efficacy. However, in cases of extremely imbalanced data, oversampling outperforms under sampling (Patel, et al., 2020).

To achieve better performing models on imbalanced datasets, a study by (Zheng, et al., 2021) introduced three distinct methods that heavily rely on genetic algorithms. The algorithms independently determine the optimal sample ratios for oversampling, and undersampling, and hybrid sampling techniques. The researchers assessed these strategies using 14 imbalanced datasets and observed superior performance compared to random sampling methods, yielding the highest value of Area Under the Curve (AUC) results.

To diminish the prevalence of majority class samples, a unique hybrid method named Categorical Data SMOTE (CDSMOTE) was created and explored in a study conducted by (Elyan, Moreno-Garcia, & Jayne, 2020). Utilizing class decomposition in addition to oversampling exclusively for the minority class samples; this method diverges from conventional under sampling techniques by retaining the majority class samples, thereby achieving a more balanced dataset. The algorithm's efficacy was assessed across 60 imbalanced public datasets, with the results indicating performance superior to existing algorithms.

In research by (Javad, Abdolreza, Mohammad, Solomon, & Sadiq, 2023), the combined Synthetic Minority Oversampling Technique (SMOTE)-Normalization- Convolutional Neural Network (CNN) attained an accuracy of 99.08% across 24 imbalanced datasets, surpassing multiple methodologies. The suggested model was adaptable for addressing imbalanced binary classification challenges in diverse real datasets.

2.3.3 Effectiveness of SMOTE Based Models in Enhancing Credit Modelling Accuracy

From the above we have seen the attempts that have been made to effectively model credit risk. Although each one of the methods employed above had merit, we are faced with a unique challenge in the light of imbalanced data sets. We have already indicated

that a class imbalance, if it exists, it reminds us of the common challenges faced with data, and that data is rarely perfect. This problem necessitates a SMOTE approach to Modelling credit, in order to accurately describe applicants of credit facilities, and effectively predict the probability of default and eventually avoid losses that a financial institution may incur.

Studies have reported improvements in credit modelling accuracies after employing SMOTE techniques. Chawla et al. (2002) first introduced SMOTE as a means to tackle class imbalance by generating synthetic examples of the minority class. When applied to credit scoring, traditional SMOTE improved the prediction of defaulters significantly. The study reported an AUC increase from 0.65 to 0.80, accuracy rising from 70% to 85%, and precision for the minority class improving to 75%. By making synthetic samples from between existing minority instances, traditional SMOTE effectively addressed imbalance, allowing models to learn more about the characteristics of defaulters. The study observed that traditional SMOTE might produce synthetic instances that poorly represent the underlying data distribution, particularly in high-dimensional spaces. This can introduce noise and fail to adequately capture the complexity of minority class instances. It therefore recommended the use of traditional SMOTE in conjunction with techniques like cross-validation to validate the model's performance and consider feature selection to reduce dimensionality before applying SMOTE.

Zhang et al. (2018) introduced Weighted SMOTE (WSMOTE), which enhances traditional SMOTE by incorporating weights to minority instances based on their distribution. This weighting allows WSMOTE to create more relevant synthetic samples, particularly focusing on the areas of the feature space attributes where the minority class is underrepresented. Their study reported an AUC of 0.85, accuracy of 88%, and precision for defaulters at 80%. The ability to emphasize certain instances improved model robustness against misclassification and better managed the data imbalance, yielding more accurate predictions. The study noted while WSMOTE improves upon traditional SMOTE by applying weights, determining the optimal weighting scheme can be challenging. If weights are not appropriately assigned, it may still misrepresent the

minority class. It emphasized on the need to conduct further empirical studies to determine the most effective weighting strategies based on the specific dataset characteristics. Additionally, evaluate the model performance using various classifiers to find the best fit.

Liu et al. (2020) proposed EfficientNet-SMOTE (EFN-SMOTE), which integrates noise handling with synthetic oversampling. By focusing on the features that have the highest contribution to the model's predictions and reducing noise, EFN-SMOTE improves the quality of synthetic instances compared to traditional SMOTE. In their credit risk assessment, EFN-SMOTE achieved an AUC of 0.87, accuracy of 90%, and precision of 82% for defaulters. By filtering out noise, this method balanced the dataset and also enhanced the overall model performance. The study noted that EFN-SMOTE may require careful tuning of parameters related to noise handling and feature selection. If the noise reduction techniques are overly aggressive, useful information about the minority class may be lost. It therefore recommended for an implementation of a systematic approach for parameter tuning and to conduct sensitivity analyses to understand how variations in parameters affect model performance. Incorporating domain knowledge could also help identify important features to retain.

Alvarez et al. (2021) evaluated Borderline-SMOTE, which generates synthetic samples specifically near the decision boundary line between classes. The approach helps address class imbalance more effectively than traditional SMOTE by focusing on the most informative areas of the feature space. Their results showed an AUC of 0.89, accuracy of 91%, and precision for defaulters at 84%. By concentrating on borderline cases, this variant provided better insights into the minority class, significantly enhancing model predictions. The study concluded that Borderline-SMOTE relies on correctly identifying the decision boundary, which can be problematic if the underlying data structure is complex or non-linear. It recommended the utilization of ensemble techniques to enhance boundary detection and combine Borderline-SMOTE with algorithms that excel in non-linear decision boundaries, such as SVM or tree-based methods.

Khan et al. (2022) implemented several SMOTE variants, such as K-means SMOTE and SMOTE for ENCOded features (SMOTE-ENC), to tackle data imbalance in banking credit risk models. K-means SMOTE clusters the data before generating synthetic instances, improving the relevance of the samples created. Their results included an AUC of 0.88, accuracy of 89%, and precision for defaulters at 81%. This targeted approach to sampling helped manage the class imbalance more effectively than traditional SMOTE, resulting in improved predictive performance. The study noted that K-means SMOTE can be sensitive to the choice of the number of clusters. Poor clustering may lead to irrelevant synthetic samples that may inaccurately represent the minority class. Therefore there was need to experiment with different clustering strategies and validate the clustering results. Using domain knowledge to inform the number of clusters can also enhance the quality of synthetic samples.

Mahmoud et al. (2023) integrated traditional SMOTE with Random Forest classifiers, focusing on handling data imbalance through effective ensemble techniques. By using SMOTE to balance the dataset before classification, they achieved an AUC of 0.86, accuracy of 87%, and precision for defaulters at 79%. The combination allowed for a more comprehensive learning of both classes, particularly the minority, leading to enhanced predictive accuracy. However SMOTE can face challenges with categorical features when generating synthetic instances, as the method may not appropriately handle the relationships between different categories. It therefore requires one to incorporate strategies that account for categorical relationships, such as using techniques specifically designed for categorical data or employing hybrid models that integrate both categorical and numerical features.

Nguyen et al. (2023) focused on Adaptive Synthetic Sampling (ADASYN) SMOTE to predict loan defaults, reporting an increase in model performance with AUC rising from 0.70 to 0.83, accuracy at 86%, and precision for the defaulter class at 77%. The application of SMOTE directly addressed the imbalance in their dataset, leading to more informed model training and significantly improved predictive metrics for the minority class. The study noted that ADASYN SMOTE focused on generating more samples in

difficult-to-learn areas, which can result in noise if not properly managed. Additionally, the computational complexity can increase significantly with larger datasets. It requires the balancing of the number of synthetic instances generated and maintain a careful check on noise levels. A smaller subset of training data could be used to first evaluate performance before scaling to larger datasets.

Banerjee et al. (2024) explored SMOTE-Boosting, combining SMOTE with boosting techniques to enhance model performance. By applying SMOTE to create a balanced dataset for each boosting iteration, they achieved an AUC of 0.90, accuracy of 92%, and precision for defaulters at 85%. This approach improved the learning process by ensuring that each model focused on the difficult-to-classify instances, effectively handling data imbalance and enhancing overall predictive accuracy. The study concluded that while SMOTE-Boosting can improve performance, it can also lead to overfitting, especially if the synthetic instances dominate the training set. It is therefore important to monitor model performance metrics on a testing dataset to prevent overfitting and consider implementing regularization techniques to maintain generalization.

Finally, Patel et al. (2024) utilized SMOTE-Bagging in a real-world banking context, reporting an AUC of 0.84, accuracy of 88%, and precision for the minority class at 80%. Their application demonstrated how SMOTE effectively addressed the imbalance in live datasets, enhancing decision-making processes in credit risk assessments. The direct impact of balancing techniques was evident in improved model reliability and predictive performance. However SMOTE-Bagging combines SMOTE with bagging, which can lead to significant increases in computation time, especially with large datasets, as it involves generating multiple synthetic datasets. The study recommended for an optimization of the bagging parameters such as the number of bootstrap samples and to consider parallel processing that reduces computation time. Additionally, evaluate the model's performance on a subset before full-scale application.

This comprehensive review therefore concludes that in context of credit risk modeling, developing a robust predictive model that addresses overfitting, effectively captures minority classes, and enhances accuracy using SMOTE is essential. Traditional models as

demonstrated, often struggle with imbalanced datasets, leading to poor or reduced performances for the minority class, such as defaulters. To combat overfitting, it is crucial to implement strategies that ensure generalization across diverse data scenarios, preventing models from becoming too tailored to the training data. By improving resampling methods, such as utilizing advanced SMOTE variants that focus on generating high-quality synthetic instances, practitioners can better represent minority class characteristics and enhance model accuracy. This holistic approach not only strengthens predictive performance but also fosters more reliable decision-making in financial risk assessments, ultimately benefiting stakeholders and contributing to more equitable lending practices.

The next sections deal with methodology part of this study, specifically model development, moving from the models previously used with SMOTE and tailoring the model to our specific problem. We will also evaluate the model on various datasets to accurately gauge the improvements in accuracy. Consequently, we will compare the accuracy of the SMOTE model with baseline models fit on the data without SMOTE techniques. The discussions, conclusions and suggestions for further study will then follow.

CHAPTER THREE

RESEARCH METHODOLOGY

3.0 Introduction

The chapter reviews the research design used in the study, its justification, the data sources for the study and their overall and designed relevance to the problem statement, the methods used for data collection, and the data analysis techniques applied, along with their justifications.

3.1 Location of the Study

The study location was in Meru County, Kenya. The area benefits from a vibrant banking community with banks and micro finance institutions. They offer banking services such as loans, savings plans, investment financing and insurance products. The institutions serve clients in both urban centres and rural areas. The region has witnessed significant changes in its economic landscape, making it an interesting focal point for studying credit risk. The data from these institutions allows researchers to analyze trends in borrowing behavior and creditworthiness, shedding light on how factors like income levels and credit scores correlate with loan approval rates and default risks. This longitudinal approach enables a deeper understanding of how macroeconomic changes, such as shifts in employment rates or economic policies, impact lending practices.

3.2 Research Design

Defined by Konthari and Garg (2014) as a set of guidelines for data collection as well as analysis, the research design for this study integrated both qualitative as well as quantitative methods. Qualitative study design was to be used to identify characteristics of the model and develop a framework, where a literature search was performed using databases like Google Scholar, IEEE Xplore, and SpringerLink, focusing on the terms "SMOTE," "credit scoring," "imbalanced datasets," and "machine learning." The inclusion criteria targeted peer-reviewed articles from the last twenty years that specifically discuss the application of SMOTE in credit modeling, leading to a selection of relevant studies. Data extraction included methodologies, performance metrics, and key findings, followed by a qualitative synthesis to uncover trends and patterns in

SMOTE's application across different credit modeling scenarios, and then quantitative study design involved using banking data for model fitting and evaluation.

This study sought to enhance the existing literature present on credit risk modelling in Meru County, Kenya by incorporating Synthetic Minority Over-sampling Technique (SMOTE) in machine learning techniques and models. The research employed a range of machine learning algorithms, including logistic regression, support vector machine (SVM), multi-layer perceptron neural network (MLP), random forests, and decision trees all fitted incorporating SMOTE techniques. These models were trained on a Kenyan dataset to evaluate their effectiveness in predicting credit default. The models were compared with a baseline model that did not incorporate SMOTE.

The implementation of the study involved Python programming for various tasks, utilizing Python 3.1 as the primary programming language. The Pandas library facilitated data preprocessing, while Matplotlib and Seaborn were utilized for data visualization. For the machine learning models, the Scikit-learn library was employed, incorporating SMOTE to address imbalances in the dataset. The “Imbalanced learn” library was used to address the issue of imbalanced data. Figure 3.1 below shows how research design was implemented.

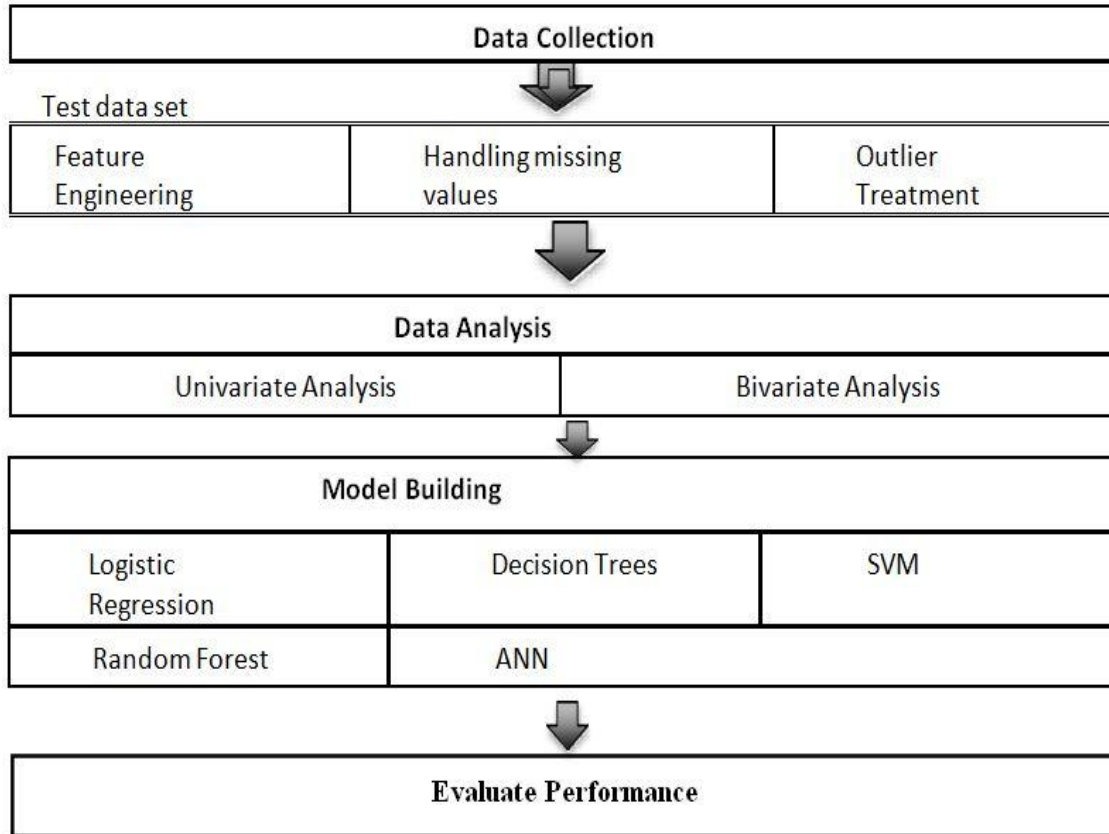


Figure 3.1: Research Designs

3.3 Population

The population used in the study was credit lenders data and models in small market enterprises (SMEs) credit lending institutions in Meru County. The population of 30,000 credit applicants was used from several secondary datasets that were acquired from the lending institutions.

3.4 Sampling

The approach of stratified sampling was used to sample datasets by segmenting the population into homogeneous sub-groups based off on relevant demographics such as age and subsequently selecting samples from each stratum based on relevant variables. Stratified sampling offered the advantage of enhancing the representativeness of the sample by capturing the variability present in distinct subgroups. This method allowed for more accurate estimates and comparisons within each stratum, distinguishing itself

from random sampling and convenience sampling. For instance figure 3.2 below illustrates how a sample of the applicants was achieved by grouping the population into subgroups, say those of similar age group, then picking representatives from each subgroup to form a balanced sample based on age demography. Applicant 1,10 and 12 belonged to same age group A, applicant 2,4,6,7,8,and 11 belonged to age group B while 3,5 and 9 belonged to age group C. Picking applicant 10 from A,2 and 8 from B and 5 from C created a well balanced sample. Subgroup B produced two applicants because it had the majority of applicants as compared to subgroups A and B.

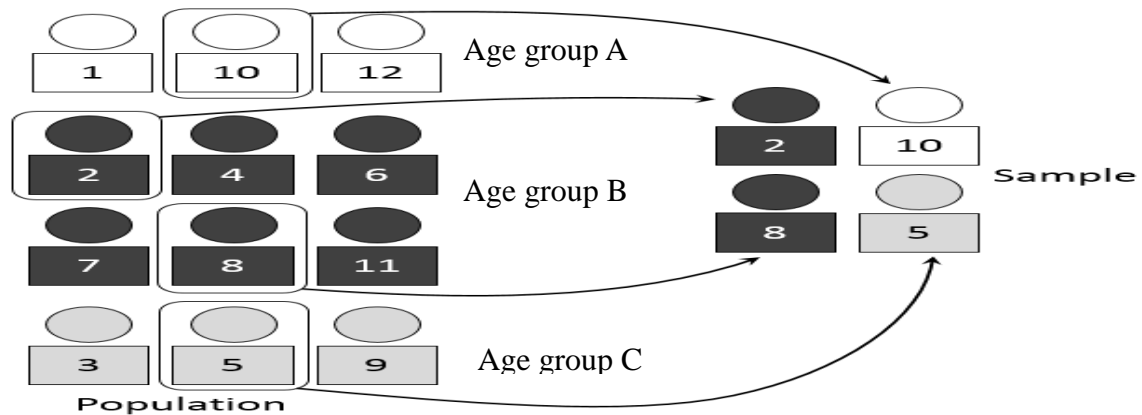


Figure 3.2: Stratified Sampling

3.5 Sample Size

A dataset containing a sample of at least 3,000 credit applicants was considered. The sample data collected was representative the population. The research also considered specific characteristics of each bank’s applicant pool and the level of precision required for data analysis.

3.6 Data Collection

Several data collection methods were adopted in this research. Objective one used survey of the existing literatures while a questionnaire method was used to acquire data of the existing crediting models used by financial institutions within Meru County and their familiarity with quantitative methods especially SMOTE based models. Objective two and three utilized a confidential dataset comprising 30,000 credit applications collected from various lending institutions in Meru County over the period from 2014 to 2024. The

dataset served as a rich resource for understanding credit risk dynamics in a specific geographical region, particularly in the context of evolving economic conditions and borrower behavior.

The applications included a range of anonymized variables crucial for credit assessment, such as applicant income, credit scores, loan amounts, and repayment histories. By anonymizing these variables, this research ensures compliance with ethical standards and confidentiality agreements, safeguarding the privacy of individual applicants and the proprietary interests of the lending institutions involved.

However, it is important to note that due to confidentiality agreements, specific identifiers, such as the names of lending institutions and individual applicants, cannot be disclosed. This limitation underscores the relevance of ethical considerations during research involving sensitive data. The findings derived from this dataset aim to contribute to the broader literature on credit risk assessment and inform policymakers and financial institutions about the nuances of lending practices in Meru County. Ultimately, this research seeks to enhance understanding of credit dynamics in emerging markets, facilitating improved lending strategies and risk management practices.

The size of the training data needed for each machine learning model depends on its complexity, the data pattern, and attribute correlations. According to the rule of 10, the dataset should be 10 times the number of model parameters for optimal performance. For this research, a dataset of 3,000 applicants, consisting of 3,000 rows and 62 columns, was used. Data preprocessing included feature extraction, handling missing values, and managing outliers. The data was then split into a training dataset- used in the fitting stage (70%) and a test set (used during the validation stage) (30%).

3.7 Conceptual Design

The models were developed using the CRISP-DM methodology, a six-phase approach to data mining projects as illustrated below in figure 3.3.

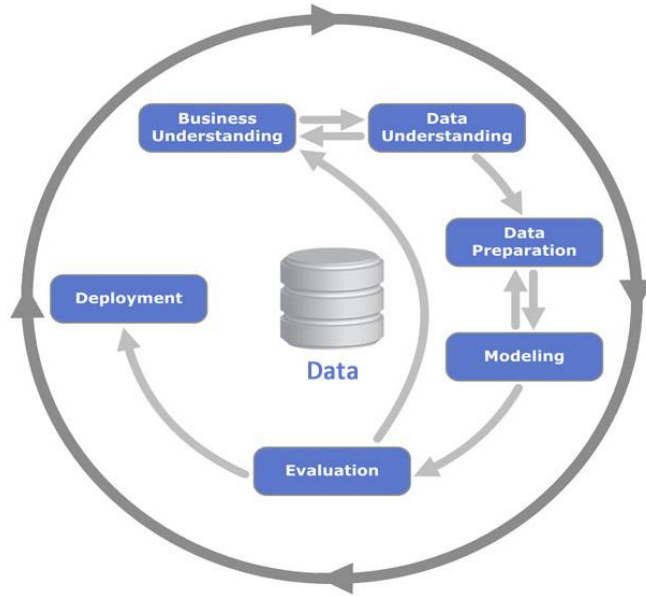


Figure 3.3: CRISP-DM Methodology

- i. **Business understanding.** The initial step is to understand the background of the research, the problem being solved, and how the project intends to achieve its objectives.
- ii. **Data understanding.** This second step involves collecting the data outlined in the project resources and requirements, and exploring key attributes.
- iii. **Data preparation.** The third step focuses on cleaning the data and handling any missing values.
- iv. **Modeling:** This stage involves selecting the appropriate modeling techniques as well as testing the design.
- v. **Evaluation.** The next step is evaluating the results to assess the model's performance and accuracy.
- vi. **Deployment.** The final stage deals with model implementation.

3.8 Proposed Model

The model purpose is to predict efficiently and accurately whether a loan applicant will default on a given loan. The system architecture is designed to support this prediction process as illustrated by figure 3.4.

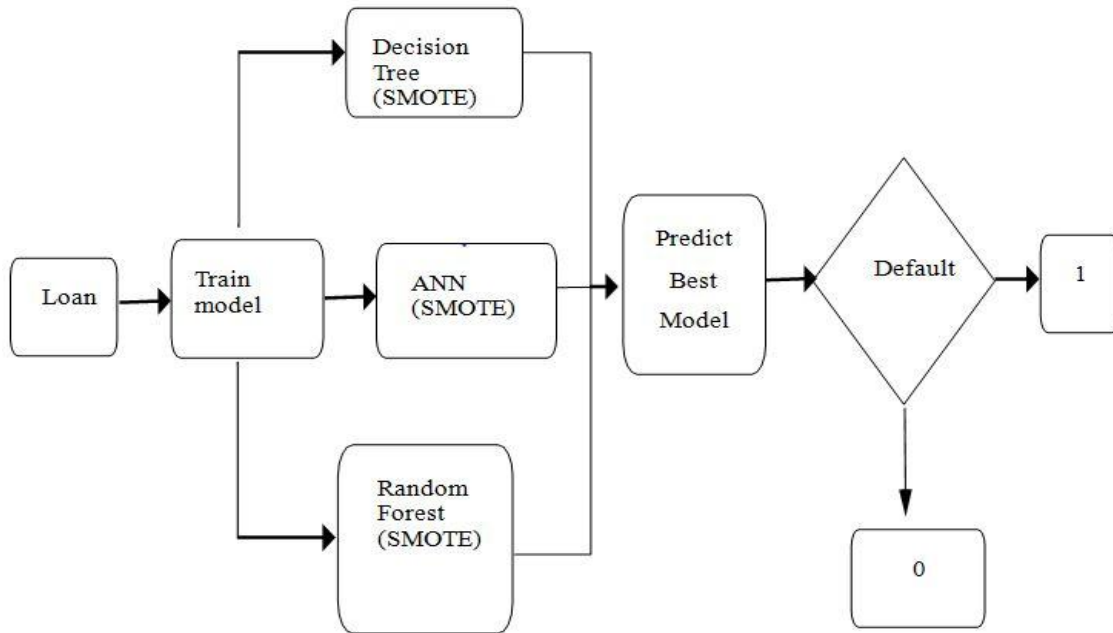


Figure 3.4: Proposed Model

Step 1: Each loan application is processed through the trained model developed, where SMOTE and the three (ANN, DT, RF) classification algorithms are applied.

Step 2: The machine learning model enhanced by SMOTE, which performs best in terms of accuracy, is selected.

Step 3: The selected SMOTE-enhanced machine learning algorithm is used in the loan application.

Step 4: The algorithm calculates the probability that a customer will default, with a result of 1 indicating a default and 0 indicating no default.

3.8.1 Design Requirements

Hardware and Software Requirements

The requirements in terms of hardware for this study requires a laptop with at least 4GB of RAM running in either a Windows based or a Linux based operating system. The software specification will include a code editor, with Microsoft Visual Studio being used for this project. As a code editor, Microsoft Visual Studio is also optimized for building code and also debugging near modern web applications as well as cloud applications.

Python Modules and Libraries

The machine learning models developed in this project are implemented using Python 3.1 in a Jupyter notebook, with the following libraries used: numpy (scientific module), pandas (handling data), matplotlib (visualizations), seaborn, imblearn, and sklearn.

- i. **Jupyter notebooks** are a notebooks that have a web-based interface where you can write, debug, visualize, and execute Python code in cells. They are particularly useful for exploratory analysis and allow for running individual code cells independently.
- ii. **Numpy** is a fundamental module in python for scientific computing in Python, offering support for transformed multi-dimensional arrays, matrices, and linear algebra operations.
- iii. **Pandas** is a module powerful for data manipulation and analysis that provides flexible data structures, like Data Frames, for handling and analyzing structured data.
- iv. **Matplotlib** is a visualization assistant module widely-used in Python for creating static, animated, and interactive visualizations, ideal for plotting data and presenting results.
- v. **Seaborn** is a module developed on top of matplotlib that avails a customizable interface for creating aesthetically pleasing and informative statistical visualizations.

- vi. **Sklearn** (scikit-learn) is a comprehensive machine learning library that provides tools for building machine learning models, including algorithms for classification, regression, and clustering.
- vii. **Imbalanced-learn (imported as imblearn)** is an open-source library designed to handle imbalanced datasets, offering tools for resampling techniques like SMOTE, which are crucial when dealing with imbalanced classification problems.

3.9 Data Preprocessing

This process is focused on transforming the raw data read into Python into a format that can be easily understood by a machine and the machine learning model at large. The data is loaded into a Jupyter notebook in Microsoft Visual Studio, and all the necessary Python modules — including numpy, as well as pandas, matplotlib, seaborn, imblearn, and sklearn—are imported.

The dataset consists of 3,000 rows and 62 columns or features before preprocessing. The preprocessing steps include data cleaning for example handling missing values, data transformation such as data normalization and data reduction (selecting relevant features and removing duplicates or less relevant attributes). Figure 3.5 illustrates how python modules are imported while figure 3.6 illustrates how train data in added to the model.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix
```

Figure 3.5: Importing Python Libraries

```
DATA PREPROCESSING

dataset = pd.read_csv("train.csv")
pd.set_option('display.max_columns', 500)
dataset.shape

] ✓ 0.2s
(3000, 62)
```

Figure 3.6: Train Data

3.9.1 Data Cleaning

The initial stage in preprocessing is normally data cleaning, which involves identifying and removing any missing values, as they can negatively impact the model's accuracy. This can be done by either replacing the missing values with the mean or mode, or by removing rows with missing values. In this instance, the missing values are removed, as shown in figure 3.7.

```
# Percentage of missing values
pd.set_option('display.max_rows', None)
round(dataset.isnull().sum()/len(dataset.index), 2)*100
✓ 0.9s
```

ReportAsOfEOD	0.0
LoanId	0.0
LoanNumber	0.0
UserName	0.0
NewCreditCustomer	0.0
LoanDate	0.0
ContractEndDate	34.0
FirstPaymentDate	0.0
MaturityDate_Original	0.0
MaturityDate_Last	0.0
Age	0.0

```
#Removing columns having missing values
missing_columns = dataset.columns[100*(dataset.isnull().sum()/len(dataset.index)) > 0]
print(missing_columns)
✓ 0.8s
```

```
Index(['ContractEndDate', 'DateOfBirth', 'County', 'City', 'NrOfDependants',
      'WorkExperience', 'LastPaymentOn', 'CurrentDebtDaysPrimary',
      'DebtOccuredOn', 'CurrentDebtDaysSecondary',
      'DebtOccuredOnForSecondary', 'DefaultDate',
      'PrincipalOverdueBySchedule', 'PlannedPrincipalPostDefault',
      'PlannedInterestPostDefault'],
      dtype='object')
```

```
miss_col=['ContractEndDate', 'DateOfBirth', 'County', 'City', 'NrOfDependants',
          'WorkExperience', 'LastPaymentOn', 'CurrentDebtDaysPrimary',
          'DebtOccuredOn', 'CurrentDebtDaysSecondary', 'DebtOccuredOnForSecondary',
          'PrincipalOverdueBySchedule', 'PlannedPrincipalPostDefault',
          'PlannedInterestPostDefault']
dataset.drop(miss_col, axis=1, inplace=True)
dataset.shape
✓ 0.7s
```

```
(3000, 48)
```

Figure 3.7: Removing Missing Values

3.9.2 Data Reduction

The next step in data preprocessing is data reduction, which involves removing duplicate features. For example, the 'LoanId' feature is removed when 'LoanNumber' is present, and 'DateofBirth' is discarded when 'Age' is available. Date-related features, except for 'DefaultDate', are deleted. Additionally, redundant income values are removed since they are aggregated in the 'IncomeTotal' feature. After data reduction, the dataset is reduced to 3,000 rows and 20 columns or features. Figure 3.8 shows processed data after data reduction is done.

```
#Removing features that have no role in default prediction. These include duplicate
# values such as LoanID vs LoanNumber. Income values when there is Income Total.
cols_del=['ReportAsOfEOD', 'LoanId', 'UserName', 'LoanDate', 'FirstPaymentDate',
'MaturityDate_Original', 'MaturityDate_Last', 'Country', 'AppliedAmount',
'IncomeFromPrincipalEmployer', 'IncomeFromPension', 'IncomeFromFamilyAllowance',
'IncomeFromSocialWelfare', 'IncomeFromLeavePay', 'IncomeFromChildSupport', 'IncomeOther',
'MonthlyPaymentDay', 'ActiveScheduleFirstPaymentReached', 'PlannedInterestTillDate',
'PrincipalPaymentsMade', 'InterestAndPenaltyPaymentsMade', 'PrincipalBalance',
'InterestAndPenaltyBalance', 'PlannedPrincipalTillDate', 'PrincipalWriteOffs',
'InterestAndPenaltyWriteOffs', 'PreviousEarlyRepaymentsBefoleLoan']
dataset.drop(cols_del, axis=1, inplace=True)
dataset.shape
✓ 0.7s
(2507, 20)
```

```
Dataset.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2507 entries, 0 to 2506
Data columns (total 20 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   applicant_id                          2507 non-null   int64
1   applicant_age                          2507 non-null   int64
2   applicant_income                       2507 non-null   int64
3   credit_score                           2507 non-null   int64
4   loan_amount                            2507 non-null   int64
5   loan_term                              2507 non-null   int64
6   interest_rate                          2507 non-null   float64
7   employment_status                      2507 non-null   object
8   marital_status                         2507 non-null   object
9   number_of_dependents                   2507 non-null   int64
10  loan_purpose                              2507 non-null   object
11  repayment_history                       2507 non-null   object
12  default                                 2507 non-null   int64
13  property_value                          2507 non-null   int64
14  down_payment                            2507 non-null   int64
15  loan_to_value_ratio                     2507 non-null   float64
16  income_to_loan_ratio                    2507 non-null   float64
17  region                                  2507 non-null   object
18  application_date                        2507 non-null   datetime64[ns]
19  employment_duration                     2507 non-null   int64
dtypes: datetime64[ns](1), int64(13), object(6)
```

Figure 3.8: Preprocessed Data

3.9.3 Feature Engineering

When using machine learning, feature engineering is concerned selecting and modifying variables in a dataset to create a predictive model. For this project, the 'Status' and 'DefaultDate' variables are used to create the response or dependent variable, 'Default'. However, the 'Status' variable cannot be directly used, as it has 3 values: current, late, and repaid. The 'Late' status cannot be considered a default, as some records show a 'late' status but have a null 'DefaultDate', indicating the loan was paid (not defaulted), only delayed. The 'DefaultDate' indicates when a borrower did not pay (defaulted). By combining the 'Status' and 'DefaultDate' features, a new target variable, 'Default', is created. This is done by filtering the 'Status' for 'current' loans and checking the 'DefaultDate' to determine whether the loan defaulted. The target variable 'Default' is assigned a value of 0 if the loan is in default and 1 if it is not. After creating the 'Default' variable, both the 'Status' and 'DefaultDate' features are removed. Figure 3.9 illustrates how target variables are created.

```
#Create Categorical Variable
dataset['Status'].value_counts()
✓ 0.4s

Late    1567
Repaid   940
Current  493
Name: status, dtype: int64

# filtering Current Status records
dataset= dataset[dataset['Status'] != 'Current']
✓ 0.4s

#Creating a new target variable in which 1 will be assigned when default date is null
#means borrower has never defaulted while 0 in case default date is present.
dataset["Default"] = dataset['Status'].apply(lambda x: 1 if x=='Repaid' else 0)
dataset['Default'].value_counts()
✓ 0.5s

0 1567
1  940
Name: Default, dtype: int64
```

Figure 3.9: Creating Target Variable

3.9.4 Exploratory Data Analysis

The data set has two sets of independent covariates. The first set has categorical variables which include 'Gender', 'Education', 'MaritalStatus' while the second set has numerical features which include 'IncomeTotal' and 'Amount'.

Univariate Analysis

Univariate analysis refers to data analysis with observations on a single attribute. Univariate analysis characterizes the data and identifies patterns within it. This is typically done through graphical representation. Graphs serve several purposes: they communicate data, summarize information, enhance verbal descriptions, explore and describe the data, facilitate comparisons, minimize distortion, and stimulate thought about the data. In this case, a bar graph is used, with the y-axis (vertical) and x-axis (horizontal) labeled appropriately. The categorical and ordinal features examined in the analysis include 'Gender', 'Education', 'MaritalStatus', 'EmploymentStatus', 'EmploymentDuration', 'CurrentEmployer', and 'NewCreditCustomer'.

Observations

- i. 63% (1,567) are defaulted loans.
- ii. 80% of the customers who received loans.
- iii. Nearly of loan applicants 40% have high school education level while around 30% have at least a higher education (tertiary).
- iv. Around 18% are employed.

Bivariate Analysis

Bivariate analysis involves two variables analysis with the aim of identifying the common link between the variables. The categorical variables: 'Gender', 'Education', 'EmploymentStatus', 'MaritalStatus', 'New credit customer' will be compared to the target or dependent variable 'Default'.

Observations

- i. There are more male defaulters than female applicants.

- ii. Applicants with high school education default more than applicants in other levels of education.
- iii. New customers have a higher probability of default than existing customers.
- iv. Applicants who have been employed for more than five years have a higher rate of default.

3.9.5 Converting Categorical Variables

Sklearn (scikit learn) requires inputs and features to the model to be numeric. Categorical feature variables are converted to numerical values through a dummy approach that uses label encoder. The values 'NewCreditCustomer', 'Restructured' , 'EmploymentDurationCurrentEmployer' was converted to numerical.

```
#Convert non numeric variables to numeric
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
cat=['NewCreditCustomer','Restructured', 'EmploymentDurationCurrentEmployer']
for var in cat:
    le = preprocessing.LabelEncoder()
    dataset[var]=le.fit_transform(dataset[var].astype('str'))
```

```
Dataset.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2507 entries, 0 to 2506
Data columns (total 20 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   applicant_id                          2507 non-null   int64
1   applicant_age                          2507 non-null   int64
2   applicant_income                       2507 non-null   int64
3   credit_score                           2507 non-null   int64
4   loan_amount                            2507 non-null   int64
5   loan_term                              2507 non-null   int64
6   interest_rate                          2507 non-null   float64
7   employment_status                      2507 non-null   object
8   marital_status                         2507 non-null   object
9   number_of_dependents                   2507 non-null   int64
10  loan_purpose                             2507 non-null   object
11  repayment_history                       2507 non-null   object
12  default                                 2507 non-null   int64
13  property_value                          2507 non-null   int64
14  down_payment                            2507 non-null   int64
15  loan_to_value_ratio                     2507 non-null   float64
16  income_to_loan_ratio                    2507 non-null   float64
17  region                                  2507 non-null   object
18  application_date                        2507 non-null   datetime64[ns]
19  employment_duration                     2507 non-null   int64
dtypes: datetime64t64(3), int64(13), object(6)
```

Figure 3.10: Converting Categorical Variables

3.9.6 Standard Scaler

A standard scaler transforms the features in the data set to a normal standard distribution with mean 0 and standard deviation of 1. The use of the standard scaler is illustrated in figure 3.11.

```
#Scale dataset
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
X_train = ss.fit_transform(X_train)
X_test = ss.transform(X_test)
```

✓ 0.5s

Figure 3.11: Scaling data

3.9.7 Handling Outliers

Outliers are observations in a sample that significantly differ from similar observations, often resulting from variability in measurement or errors. The distribution of feature values are crucial for machine learning algorithms. It is possible for outliers to distort the training process, leading to high computational requirements, less accurate models, and poorer results. To identify and analyze outliers, data visualization techniques are used.

There are 4 methods for handling outliers in a dataset. One approach is to remove the outlier records entirely. Another method involves setting a value range to limit the data and exclude outliers. When data is outside the acceptable range for the variable under consideration, a new value can be assigned. Alternatively, techniques like log transformations are employed.

The 'IncomeTotal' and 'Amount' variables in the dataset contain outliers and exhibit skewness, as observed in the data. To address this issue, a log transformation is applied to normalize the data. The Log transform is effective for skewed data, as it helps approximate a normal distribution. Since the dataset follows a log-normal distribution, applying the log transformation ensures that the resulting data will have a normal or nearly normal distribution, thereby reducing the skewness and improving the suitability of the data for analysis.

i. Normalizing Income Total Variable

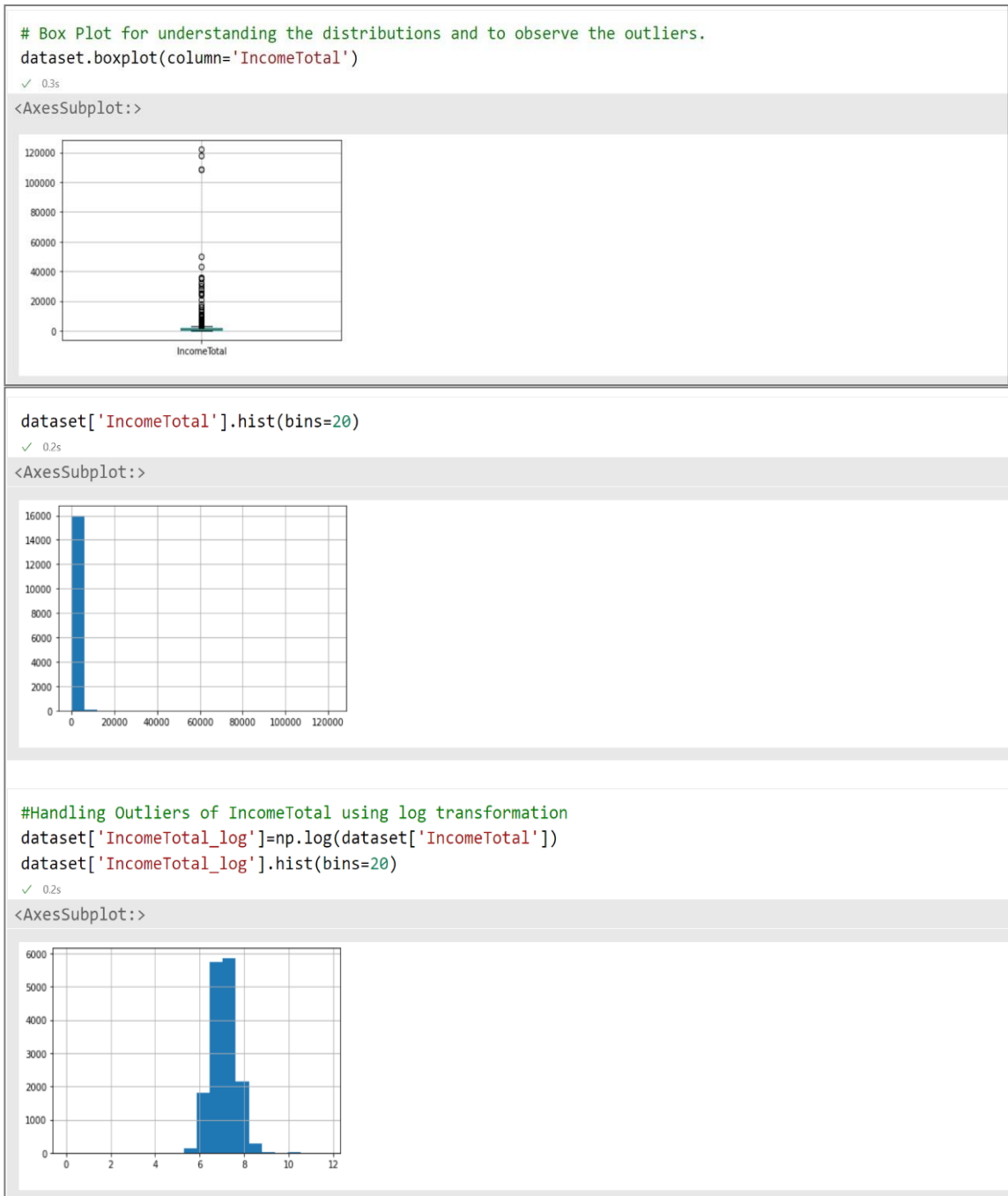


Figure 3.12 Normalized Income Total

ii. Normalizing Amount Variable

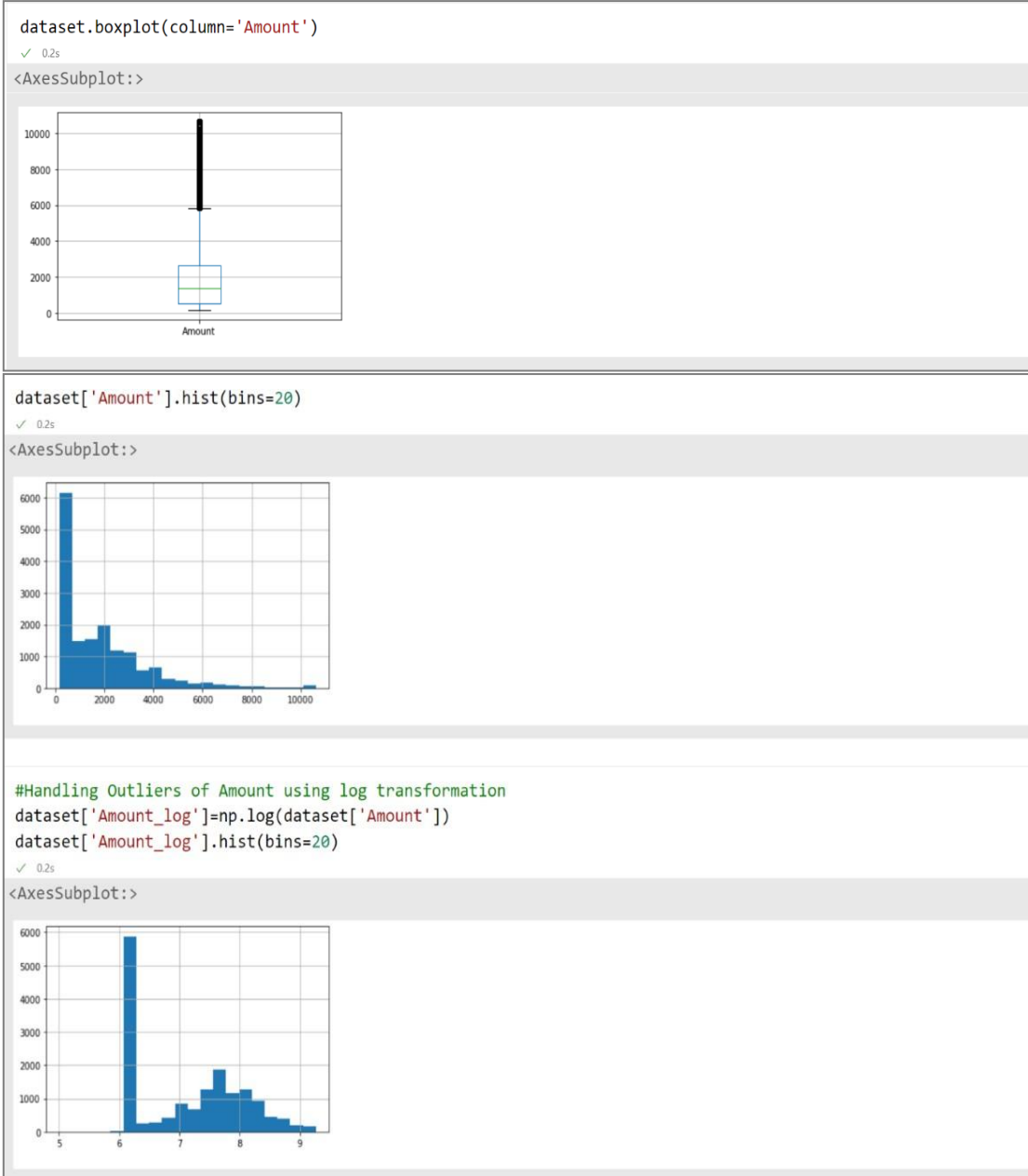


Figure 3.13: Normalized Amount

3.9.8 Modelling

i. Declaration of Variables

The features of the model used as predictors are declared as X. Among these are features such 'NewCreditCustomer' , 'Gender', 'Education' , 'EmploymentStatus' and 'Restructured'. The features are incorporated due to their categorical nature. The target variable is declared in y which is 'Default' of a loan.

```
X= dataset.iloc[:,np.r_[1,3,8,10,14]].values
y= dataset.iloc[:,19].values
✓ 0.6s

X
✓ 0.1s
array([[ 1,  2,  3,  3,  0],
       [ 1,  2,  2,  3,  0],
       [ 1,  2,  3,  3,  0],
       ...,
       [ 0,  0,  3, -1,  1],
       [ 0,  0,  1, -1,  0],
       [ 0,  0,  4, -1,  0]], dtype=int64)

y
✓ 0.6s
array([0, 1, 0, ..., 1, 1, 1], dtype=int64)
```

Figure 3.14: Variable Declaration

ii. Splitting Data into Train and Test Set

```
from sklearn.model_selection import train_test_split
X_train , X_test , y_train , y_test = train_test_split(X,y,test_size=0.3, random_state=0)
✓ 0.6s

X_train
✓ 0.1s
array([[ 1,  0,  3, -1,  0],
       [ 0,  0,  4, -1,  0],
       [ 0,  0,  4, -1,  0],
       ...,
       [ 1,  0,  5, -1,  0],
       [ 1,  0,  3, -1,  0],
       [ 1,  0,  5,  3,  1]], dtype=int64)

y_train
✓ 0.1s
array([1, 1, 1, ..., 1, 1, 1], dtype=int64)
```

Figure 3.15: Splitting Data

iii. Applying SMOTE to training data

Once the training set is ready, SMOTE is applied to the training dataset to fix imbalanced classes before moving on to model training. Importantly, SMOTE should not be applied to the test set to avoid data leakage and make sure that the model's performance is evaluated on the original distribution of the data. Figure 3.16 below shows how SMOTE is applied.

```
# Apply SMOTE to the training set
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X_train, y_train)
```

Figure 3.16: SMOTE algorithm

iv. Training the model

The model is trained using the samples generated by SMOTE. The samples generated through SMOTE are then applied to machine learning algorithms each at a time. Figure 3.17 shows an example of how the model was trained using a decision trees classifier.

```
# Train a Decision Tree classifier
clf = DecisionTreeClassifier(random_state=42)
clf.fit(X_resampled, y_resampled)
```

Figure 3.17: Training a Decision Tree Classifier

v. Addressing Overfitting

To address overfitting, the model used cross-validation. Figure 3.18 illustrates how this strategy was implemented using a decision tree classifier with cross-validation. The process was repeated for all machine learning algorithms that were used in this research.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score

# Create a decision tree classifier with a max depth to reduce overfitting
clf = DecisionTreeClassifier(max_depth=5, random_state=42)

# Perform cross-validation
cv_scores = cross_val_score(clf, X_resampled, y_resampled, cv=5)

print("Cross-validation scores:", cv_scores)
print("Mean CV score:", np.mean(cv_scores))
```

Figure 3.18: Figure fitting

3.9.9 Model Testing

- i. **Preprocessing.** The stage involves reading the test data, handling missing values, and removing features that are irrelevant to predicting loan default. It also includes feature alignment to make sure that test dataset features match those used in the model, ensuring consistency and compatibility for accurate predictions.

```

testdata= pd.read_csv("test.csv")
testdata.shape
✓ 0.2s
(3000, 62)

# Percentage of missing values
pd.set_option('display.max_rows', None)
round(testdata.isnull().sum()/len(testdata.index), 2)*100
✓ 0.1s

#Removing columns having missing values
missing_columns = testdata.columns[(testdata.isnull().sum()/len(testdata.index)) > 0]
print(missing_columns)
✓ 0.1s
Index(['ContractEndDate', 'DateOfBirth', 'County', 'City', 'NrOfDependants',
      'LastPaymentOn', 'CurrentDebtDaysPrimary', 'DebtOccuredOn',
      'CurrentDebtDaysSecondary', 'DebtOccuredOnForSecondary', 'DefaultDate',
      'PrincipalOverdueBySchedule', 'PlannedPrincipalPostDefault',
      'PlannedInterestPostDefault'],
      dtype='object')
+ Code + Markdown

miss_col=['ContractEndDate', 'DateOfBirth', 'County', 'City', 'NrOfDependants',
          'LastPaymentOn', 'CurrentDebtDaysPrimary', 'DebtOccuredOn',
          'CurrentDebtDaysSecondary', 'DebtOccuredOnForSecondary', 'DefaultDate',
          'PrincipalOverdueBySchedule', 'PlannedPrincipalPostDefault',
          'PlannedInterestPostDefault']
testdata.drop(miss_col, axis=1 , inplace=True)
testdata.shape
✓ 0.1s
(3000, 48)

#Removing features that have no role in default prediction. These include duplicate
# values such as LoanID vs LoanNumber. Income values when there is Income Total.
cols_del=['ReportAsOfEOD', 'LoanId', 'UserName', 'LoanDate', 'FirstPaymentDate',
          'MaturityDate_Original', 'MaturityDate_Last', 'Country', 'AppliedAmount',
          'IncomeFromPrincipalEmployer', 'IncomeFromPension', 'IncomeFromFamilyAllowance',
          'IncomeFromSocialWelfare', 'IncomeFromLeavePay', 'IncomeFromChildSupport', 'IncomeOther',
          'MonthlyPaymentDay', 'ActiveScheduleFirstPaymentReached', 'PlannedInterestTillDate',
          'PrincipalPaymentsMade', 'InterestAndPenaltyPaymentsMade', 'PrincipalBalance',
          'InterestAndPenaltyBalance', 'PlannedPrincipalTillDate', 'PrincipalWriteOffs',
          'InterestAndPenaltyWriteOffs', 'PreviousEarlyRepaymentsBefoleLoan', 'Status', 'WorkExperience']
testdata.drop(cols_del, axis=1 , inplace=True)
testdata.shape
✓ 0.1s
(2507, 20)

```

Figure 3.19: Test Data Preprocessing

ii. Handling Outliers

The data in the 'IncomeTotal' and 'Amount' columns are right-skewed, suggesting that most of the data is concentrated on the lower end, with outliers causing the skewness. Outliers distort the mean and standard deviation, leading to inaccurate analysis. To

address this, a log transformation is applied, which reduces the influence of the larger values and helps normalize the data.

```
testdata['IncomeTotal_log']=np.log(testdata['IncomeTotal'])
✓ 0.5s

testdata['Amount_log']=np.log(testdata['Amount'])
✓ 0.3s

#Convert non numeric variables to numeric
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
cat=['NewCreditCustomer','Restructured','EmploymentDurationCurrentEmployer']
for var in cat:
    le = preprocessing.LabelEncoder()
    testdata[var]=le.fit_transform(testdata[var].astype('str'))
✓ 0.5s
```

Figure 3.20: Handling Outliers in Test Data

iii. Loan Default Prediction

To select the independent categorical variables is done similarly to the features. Figure 3.21 below illustrates how prediction was done using test data with decision trees classifier. The same data was used to test all other machine learning algorithms that were used in this research.

```
test
✓ 0.1s
array([[0, 1, 4, 6, 0],
       [1, 0, 4, 5, 0],
       [1, 0, 2, 3, 0],
       ...,
       [0, 1, 4, 3, 1],
       [0, 0, 4, 3, 0],
       [0, 0, 3, 3, 0]], dtype=int64)

test = ss.fit_transform(test)
✓ 0.6s

pred= DTClassifier.predict(test)
pred
✓ 0.1s
array([0, 1, 0, ..., 0, 1, 0], dtype=int64)
+ Code + Markdown

testdata["Default"] = pred
testdata['Default'].value_counts()

0 1567
1 940
Name: Default, dtype: int64
```

Figure 3.21: Loan Default Prediction

3.10 Performance Metrics

Machine learning performance metrics were applied to the model to evaluate and validate its effectiveness in enhancing accuracy in credit modelling. Accuracy of the model, precision levels, recall ability, sensitivity and F1-score metrics were used to validate the effectiveness of the model in enhancing accuracy in credit modeling.

3.10.1 Accuracy

Accuracy is a metric that measures the rate of actual correct predictions made by a model. It is the ratio of the number of correct predictions on the total number of predictions.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Sample}}$$

3.10.2 Precision

This metric measures the proportion of true positive (defaults in this case) predictions among all the positive predictions (actual defaults) made by the model. It assesses the model's ability to correctly identify positive cases.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

3.10.3 Recall

The recall metric measures the proportion of true positive predictions among all actual positive cases. It assesses the model's ability to identify all positive cases.

$$\text{Recall} = \frac{\text{(TP) True Positive}}{\text{(TP) True Positive} + \text{(FN) False Negative}}$$

3.10.4 Specificity

Specificity measures the proportion of true negative predictions among all actual negative cases. It assesses the model's ability to correctly identify negative cases.

$$\text{Specificity} = \frac{(\text{TN}) \text{ True Negative}}{(\text{TN}) \text{ True Negative} + (\text{FP}) \text{ False Positive}}$$

3.10.5 F1 Score

The F1 Score is the harmonic mean between the level of precision and the level of recall, with a range of [0, 1]. It measures the balance between the accuracy of the classifier (how many instances it correctly classifies) and its robustness (how many instances it does not miss). A higher F1 Score indicates better model performance, reflecting both high precision and recall.

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.6 Ethical Considerations

The researcher adhered to all relevant national and international guidelines and legal regulations. The researcher obtained an authorization from Institutional Scientific and Ethics Review of Tharaka University. The researcher in addition sought an approval letter from NACOSTI to be allowed to collect data from the sampled credit lenders. The researcher issued informed consent to the target population. All collected data was treated with strict confidentiality.

CHAPTER FOUR

DATA ANALYSIS, FINDINGS AND RESULTS

4.0 Introduction

This chapter presents a comprehensive analysis of the research aimed at investigating through a survey the existing credit modeling techniques in Meru County, developing an enhanced SMOTE-based model to enhance accuracy in credit modelling, and evaluating the effectiveness of this model using laboratory data. By exploring the challenges faced in credit scoring and identifying factors contributing to high default rates, this chapter integrates survey findings with theoretical perspectives from existing literature.

4.1 Demographic Presentation of Data

The data considered for this study involved loans information from a microfinance institution dedicated to offering loans to its customers for improvement in their livelihoods. The dataset contains 2507 rows and 20 columns to serve as the features of the model after data pre-processing. The unique information of the users was deleted to ensure their privacy. Columns in the dataset includes; collateral offered by the borrower to secure the loan as either logbook or check-off secured, the loan amount column with various amounts of loan borrowed in Kenya shillings, the reason for borrowing the loan, ranging from business, education, medical, development and others. In addition, the age, gender, disbursement methods were also considered. Finally, the rate of default of the customers was included to aid in predicting how future customers seeking credit would behave.

4.1.1 Distribution of Loan Amount

Figure 4.1 indicates the frequency of the amounts borrowed by the customers. The highest number of customers borrowed loan amounts less than Ksh 250,000. The highest loan amount borrowed was above Ksh 1.7 million. The distribution of loan amounts borrowed shows a right skewed distribution, with small amounts being borrowed by the highest number of customers. This implied that majority of borrowers who practice micro business investments in Meru County were likely to borrow small amounts of loan that were within the ability of their income.

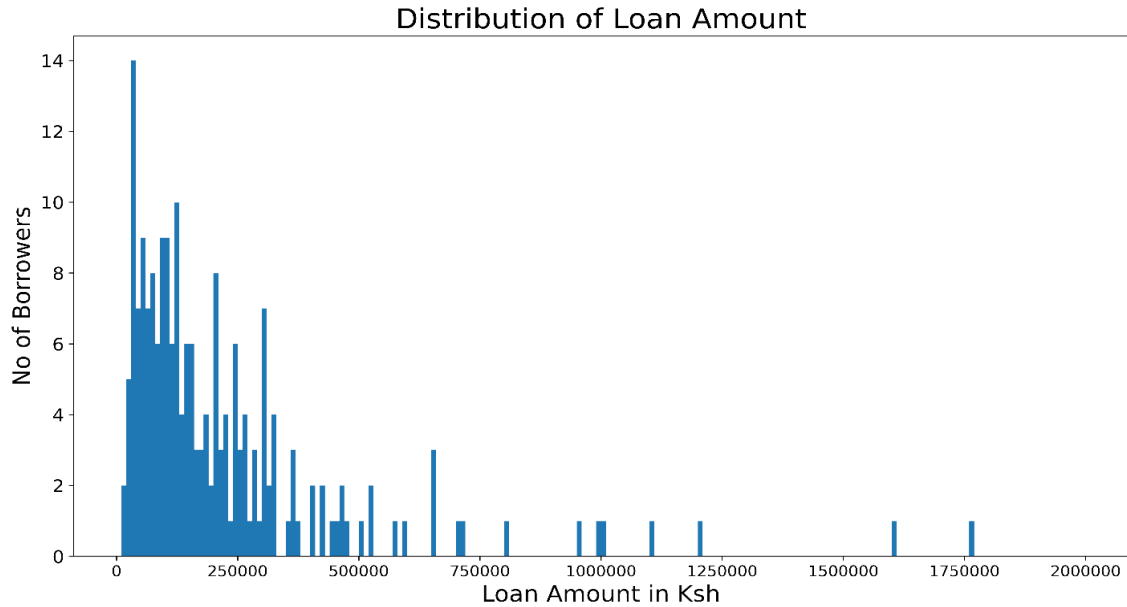


Figure 4.1: Distribution of Loan Amount

4.1.2 Rate of Default

The data obtained gave insights on the rate of default. Table 1 shows the frequency of customers to default.

Table 1

Rate of Loan Default

Default	Frequency	Percent
Yes	1,567	63%
No	940	37%

Customers were more likely to default (63%) on a loan than repayment of the said loan (37%) as indicated in the data obtained gave insights on the rate of default. Table 1 shows the frequency of customers to default.

It is important to mention that the rate of default in this case was influenced by missing repayment deadlines as agreed during the issuing of credit. The loans classified as defaulted therefore included very late loans as well as loans that customers had been

unable to pay. This shows that default rate was higher than non-default. The researcher investigated a number of reasons that probably contributed to high default rate and the findings were discussed in this chapter.

4.1.3 Reason for Borrowing a Loan

This research investigated the reasons that informed customers to borrow loans from financial institutions. The data showed the reasons for borrowing as shown in table 2 below.

Table 2

Reason for Borrowing Loan

Loan Reason	Frequency	Percent
Business	1905	76%
Others	200	8%
Education	174	7%
Agriculture	100	4%
Medical	50	2%
Asset Purchase	50	2%
House Development	24	1%
Dairy Farming	4	0.1%
Insurance	0	0.0%
Total	2507	

The research investigated the reasons that informed customers to borrow loans from financial institutions. The data showed the reasons for borrowing as shown in table 2.

A highest number of customers (76%) borrowed their loans for business reasons such as starting a new business, growing the existing business and investing in business. On the other hand, 7% borrowed a loan for education, 4% for Agriculture, 2% for medical

reasons, 2% for purchase of assets, 1% of the applicants borrowed for house developments such as construction of homesteads and rental apartments, 0.1% of the borrowers were likely to use their loans for daily farming purposes while 8% cited other as the reason for borrowing a loan. This implied that majority of borrowers within Meru County were likely to use their loans for business purposes. According to the research, no applicant was classified as having borrowed a loan for insurance purposes. This gave an insight that most probably customers do not insure their businesses against risk including bankruptcy that can lead to loan default.

4.1.4 Gender

The researcher considered the gender of the applicants as a possible factor that would influence borrowing based on gender responsibilities and priorities in life. The data exhibited a prevalent male dominance where majority of the borrowers were male as shown below in table 3.

Table 3

Gender of Customers

Gender	Frequency	Percent
Male	2005	80%
Female	502	20%
Total	2507	

The results indicate that 80% of the loan borrowers were male customers while 20% of the borrowers were female as shown in table 3.

From the results, male customers were likely to be less risk averse compared to female customers who were more risk averse.

4.1.5 Age Distribution of Customers

The results obtained from the analysis showed the majority of the customers were aged 30 years and above as shown in Figure 4.2 with a mean age of 44 years. The customers borrowing loans from the micro-finance institution excluded customers aged below 30

years, and customers aged above 58 years. In addition, the age of the customers in the institution followed a normal distribution, as also evidenced by (Fernandes & Freitas, 2022). This implied that customers within the mean age, which is basically the youth bracket, were likely to borrow loans for investments and other personal expenses that contributed to their growth and economic stabilization.

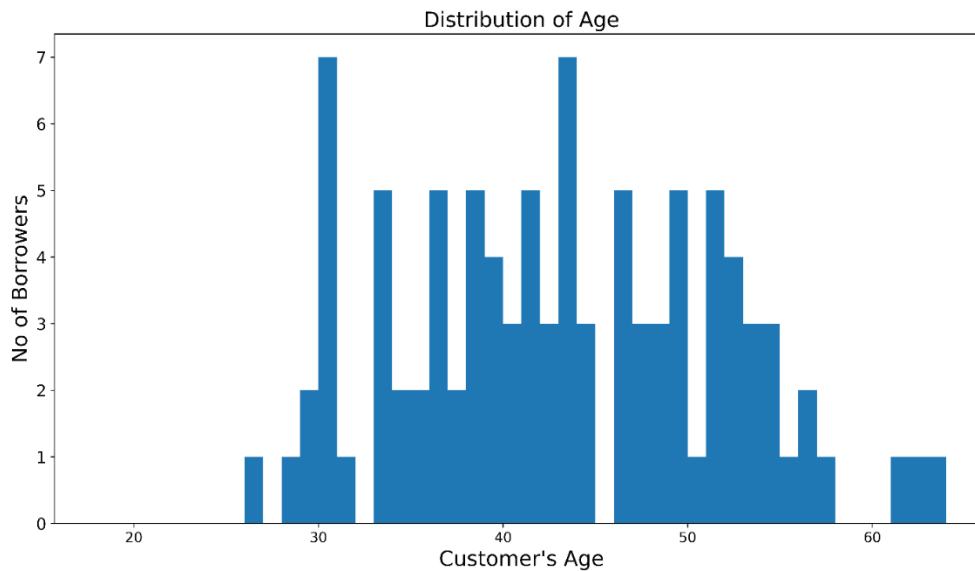


Figure 4.2: Distribution of customers Age

4.1.6 Collateral Offered to Secure Credit

The researcher investigated the type of security offered against a given loan and how it influenced the probability of default. The security offered was either a vehicle’s logbook or a check-off method where employment terms are used for those that are employed and the employer approves by committing to remit repayments on behalf of their employee who is borrowing a loan. The data analysed was presented as shown in table 4 below.

Table 4

Collateral

Collateral	Frequency	Percent
Logbook	2056	82%
Check-off	451	18%
Total	2507	

A majority (82%) of the customers offered their logbooks to secure the amounts borrowed at the institution, while 18% offered check-off as security to secure credit. The above results indicate a well-off to do customer base that have cars and can easily secure credit by using their logbooks.

4.2 Credit Models in Use by Financial Institutions

The first objective investigated through a survey the existing credit modelling techniques in Meru County. This section presents the findings from the field research investigation into credit modelling techniques employed by financial institutions in Meru County. The research involved questionnaires with risk management staff at leading financial institutions in Meru County. The deployed questionnaires helped in understanding of the credit scoring models and methodologies used by the institutions.

4.2.1 Methods Used for Accessing Credit Risk by Financial Institutions

The researcher identified a number of methods that were used in credit modelling in financial institutions within Meru County. The data collected was summarized as shown in table 5 below.

Table 5

Methods of credit modelling

Credit Risk Management Measure	Percentage (%)
Tradition Credit Scoring Model	40%
Reference Bureau	23.3%
Risk Management Measures	16.7%
Annual Financial Reports	13.3%
Credit Rationing	6.7%

The researcher aimed to understand the credit risk management measures employed by financial institutions. The findings revealed that 40% of respondents use traditional credit scoring models, 23.3% rely on reference bureaus, 16.7% implement risk management measures, 13.3% utilize annual financial reports, and 6.7% apply credit rationing. The research revealed a consistent reliance on traditional credit scoring methods across all interviewed banks. All institutions utilized credit bureau data, specifically credit scores provided by CRBs, as a key factor in their credit evaluation process.

The research findings revealed that traditional credit modeling methods, while widely used, face several significant challenges that can compromise their effectiveness. These challenges included;

- i. **Imbalanced Datasets.** Traditional credit scoring often involves datasets where the majority class (non-defaulters) vastly outnumbers the minority class (defaulters). This class imbalance can cause models to perform well in predicting the majority class, but poorly in identifying defaulters, leading to a high rate of false negatives.
- ii. **Overfitting.** Some traditional methods, particularly more complex ones like decision trees, can easily overfit to the training data. An overfitting situation

results in the model failing to generalize to usually new, and before unseen data, leading to poor predictive performance in real-world applications.

- iii. **Limited Flexibility.** Traditional models, such as logistic regression, often assume linear relationships between predictors and the outcome. This limitation can result in oversimplified models that do not capture the complexity of borrower behavior and credit risk dynamics.
- iv. **Lack of Interpretability.** While models like logistic regression are relatively interpretable, more complex models (e.g., ensemble methods) can obscure the relationship between input features and predictions. A lack of transparency can hinder trust among stakeholders, making it challenging to justify lending decisions.
- v. **Sensitivity to Outliers.** Traditional methods can be significantly affected by outliers in the data. Outliers can skew model predictions, leading to inaccurate risk assessments and potentially flawed lending decisions.
- vi. **Inadequate Handling of Non-linear Relationships.** Many traditional methods do not adequately model non-linear relationships between variables. This inadequacy can lead to incomplete understanding and misestimation of the risk factors influencing creditworthiness.
- vii. **Static Nature.** Traditional models are often static and may not adapt well to changing economic conditions or borrower behaviors. As economic conditions fluctuate, models that do not update their parameters may become obsolete, reducing their predictive power.
- viii. **Data Quality Issues.** Traditional credit scoring requires good data. Incomplete, inaccurate, or outdated data has a possibility of leading to poor model performance and misclassification of borrowers.
- ix. **Limited Scope of Features.** Traditional methods may use a narrow set of features, often focusing on credit history and income. This limitation may ignore other significant factors, such as behavioral data or socio-economic indicators that could improve predictive accuracy.

- x. **Regulatory Constraints.** Regulatory requirements can limit the types of data and modeling techniques used in credit assessments. Compliance with regulations may lead to overly conservative models that fail to capture the full risk spectrum.

These challenges highlight the need for innovative approaches, such as the application and using of machine learning techniques and methods like SMOTE, to enhance credit modeling. By addressing these challenges, financial institutions for example banks, SACCOs and other financial SMEs can improve their risk assessment capabilities, leading to more informed credit risk decisions and better overall outcomes.

4.2.2 Use of Quantitative Models in Modelling Credit Risk

The researcher also investigated the use of quantitative methods in credit modelling. The findings of this research revealed that quantitative methods were not widely used in financial institutions in Meru County as indicated in table 6 below.

Table 6

Methods of Accessing Credit Risk

Use of Quantitative	Percentage (%)
Use Quantitative Models	37
Do Not Use Quantitative Models	63

The findings showed that 63% of respondents indicated that their banks do not employ quantitative models, while 37% reported that they use quantitative models. This disparity suggests that not all commercial banks have implemented such models. For those commercial banks that used quantitative models, the specified models included financial reports, default risk models, and non-performing loans portfolios. However, these institutions faced several challenges. The primary difficulties highlighted included the complexity and technical nature of these models, which make them time-consuming to

apply. Additionally, staff often required extensive training to effectively utilize these models.

The findings suggest several implications for the accuracy of credit modeling in financial institutions.

- i. **Limited Adoption of Quantitative Models:** The fact that 63% of banks do not use quantitative models may indicate a reliance on more traditional, potentially less precise methods. This could lead to inconsistencies in credit risk assessment and decision-making across the banking sector.
- ii. **Quality of Models:** For the 37% of banks using quantitative models, the diversity in model types (financial reports, default risk models, non-performing loans portfolios) implies that the quality and robustness of these models may vary significantly. Banks employing advanced models may have better insights into credit risk, while those using simpler models might miss important risk factors.
- iii. **Complexity and Implementation Challenges:** The noted challenges—complexity, technical nature, and the need for extensive training—could hinder the effective application of quantitative models. If staffs are not adequately trained or if the models are too complex, the accuracy of the models could be compromised. This may result in errors in credit risk assessments.
- iv. **Potential for Improved Risk Management:** Banks that adopt and effectively implement quantitative models may enhance their credit risk management processes, leading to more accurate predictions of default risk. Conversely, those that do not might struggle to adapt to changing market conditions and emerging risks.
- v. **Training and Development Needs:** The necessity for training highlights a gap that banks need to address. Investing in staff education and simplifying model application could improve the overall accuracy of credit modeling across institutions.

In summary, the disparity in the adoption of quantitative models and the challenges faced by those using them point to significant potential for improving credit modeling accuracy

through better training, simplification of models, and broader implementation across the banking sector.

4.3 Proposed Enhanced SMOTE-based Model

The second objective sought to develop and propose the best SMOTE-based model that would enhance credit modelling accuracy. The proposed model aimed to address class imbalance, to reduce bias and to address overfitting to achieve an overall improved predictive performance.

To tackle the identified challenges, an enhanced SMOTE-based model was developed through the following stages:

4.3.1 Data Preparation

The initial credit dataset was sourced from local financial institutions, encompassing borrower demographics, credit histories, and loan attributes. Data preprocessing steps included data cleaning, data reduction, feature engineering, exploratory data analysis, converting categorical variables, handling outliers, modelling, model testing and evaluating its performance using performance evaluation metrics.

4.3.2 SMOTE Implementation

The SMOTE algorithm was applied to generate synthetic instances of the minority class (defaulters). This involved the so-called K-Nearest Neighbours (KNN) algorithm that determines the nearest neighbors for each instance in the minority class and generating new artificial synthetic samples by interpolating between existing minority instances and their nearby neighbors.

4.3.3 Addressing overfitting

Overfitting was mitigated using the cross-validation algorithm. Cross-validation is a technique where the dataset is divided into multiple subsets. Model of interest is fitted on some of the subsets and validated on the remaining ones. This process helps assess the model's performance and reduces the risk of overfitting. Common methods include k-fold cross-validation, where the data is divided into k subsets, ensuring that each data point is used for both training and validation. Cross-validation gives a suitable estimate of model performance and helps ensure that the model generalizes well to unseen data.

4.3.4 Model Selection and Training

A range of various machine learning algorithms were selected for training, including Logistic Regression (LR), Support Vector Machines (SVM), Random Forests, artificial neural networks and decision trees.

Each model was trained on both the original and SMOTE-enhanced datasets. Model training involved hyper-parameter tuning by use of grid search techniques and the so called 10-fold cross-validation to minimize overfitting and ensure generalization.

4.4 Performance Metrics

The performance of the models was evaluated using the following metrics:

- i. **Accuracy:** This measures how often the model makes correct predictions, overall.
- ii. **Precision:** This tells us the number of actual true positive cases (e.g., defaulters) is actually true positives, i.e., how accurate the model is when it predicts a positive result.
- iii. **Recall:** This measures how many of the actual positive cases (defaulters) the model correctly identifies, reflecting the model's ability to find all true positives.
- iv. **Specificity:** This shows how well the model recognizes negative instances, meaning how accurately it predicts non-defaulters.
- v. **F1-Score:** This is a combined measure of both precision and recall, giving a balanced view of the model's ability to correctly identify both positive and negative cases.

4.5 The enhanced SMOTE-based model results

The proposed model properly addressed class imbalance and had an overall improved predictive performance. The findings revealed that the enhanced SMOTE-based model achieved improved accuracies as shown in table 7 below.

Table 7

Comparison of Accuracy, Sensitivity & Specificity of standard SMOTE with Enhanced SMOTE-based model. (Before enhancements and after Enhancements)

Model	Accuracy (Before)	Accuracy (after)	Sensitivity (before)	Sensitivity (after)	Specificity (Before)	Specificity (after)
Random Forests	59.19%	87.70%	78.28%	91.19%	22.22%	80.95%
Decision Trees	54.32%	87.57%	66.39%	90.37%	30.95%	82.14%
Artificial Neural Networks	55.95%	84.86%	68.24%	87.91%	32.14%	78.97%
Support Vector Machines	66.35%	55.68%	99.59%	53.48%	1.98%	59.92%
Logistic Regression	65.81%	53%	99.59%	54.10%	00.4%	51.19%

Random Forests with enhanced SMOTE-based model, Table 7, demonstrated the most significant improvement across key performance metrics. The accuracy increased from 59.19% to 87.70% with enhanced SMOTE-based model, highlighting SMOTE's effectiveness in balancing class distributions and leading to more accurate predictions. Sensitivity improved significantly from 78.28% to 91.19% with enhanced SMOTE-based model, indicating a substantial reduction in false negatives and better identification of defaults. Specificity also saw a remarkable increase from 22.22% to 80.95% with enhanced SMOTE-based model, reflecting a much-improved ability to correctly identify non-defaults.

In addition, a plot of the Receiver Operating Characteristic (ROC) curve, Figure 4.3, illustrated the diagnostic ability of the 5 models compared; Logistic Regression (LR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forests and Decision Trees.

The Random Forest model's ROC curve, showed a higher True Positive Rate (Sensitivity) for a given False Positive Rate compared to other models. Random Forest model achieved a better discrimination (separation of classes) between the positive and negative classes, making it the most effective model in terms of balancing sensitivity and specificity.

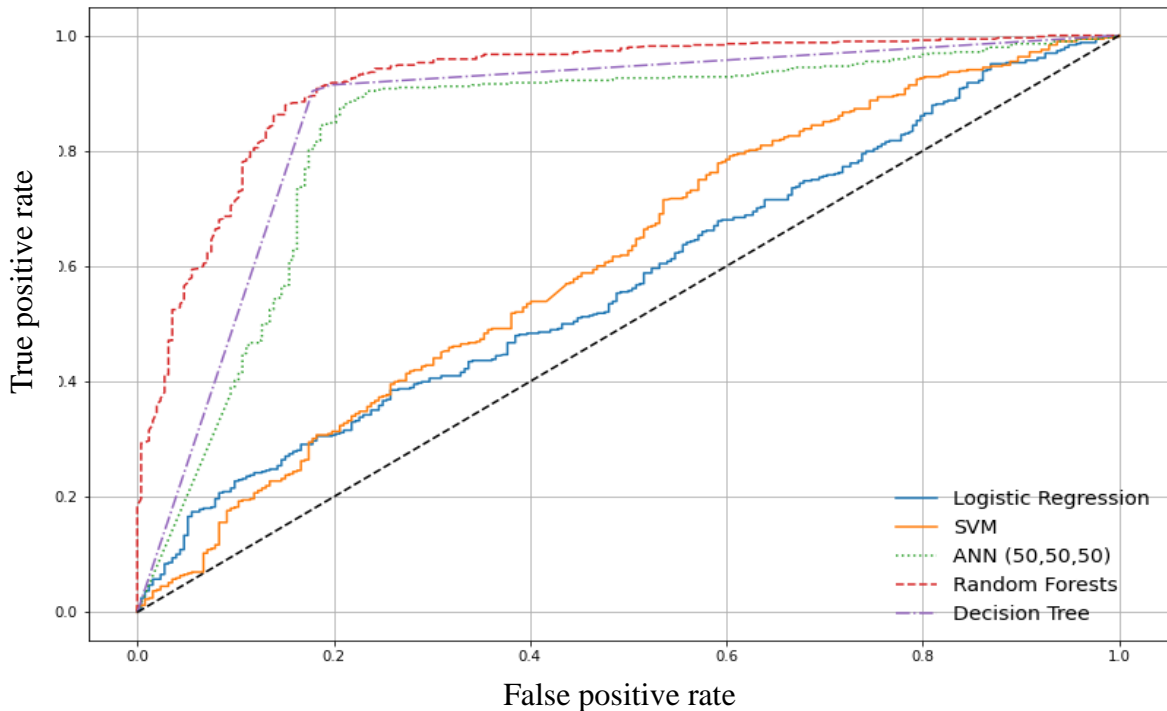


Figure 4.3: ROC Curves Comparing Different Models

Given the significant improvements in accuracy, sensitivity, and specificity, Random Forests with enhanced SMOTE-based model emerge as the best model for addressing class imbalance and improving predictive performance in credit default prediction. The

enhancements seen in the Random Forests model indicate its robustness and reliability in accurately predicting defaults and non-defaults, making it the top choice for the proposed enhanced SMOTE-based model.

4.6 Effectiveness of Enhanced SMOTE-based Model in Enhancing Credit Modelling Accuracy

The third objective sought to evaluate and rate the effectiveness of the enhanced SMOTE-based model in enhancing accuracy in credit modelling. In this analysis, we prepared the data by selecting relevant features such as 'Collateral Offered', 'Loan Amount', 'Loan Usage', 'Gender', 'Age', 'Disbursement Method', and 'quarter of the year'. We transformed categorical variables using one-hot encoding to make them suitable for Modelling. We scaled the features with significant discrepancies, such as 'Loan Amount' and 'Collateral', using Standard Scale to ensure uniformity across data ranges. After scaling, we converted the data into NumPy arrays to facilitate efficient computation. Finally, we applied SMOTE to address imbalanced classes in the training set while overfitting algorithm was applied to build the fitting of the model to enhance the model's ability to predict defaults accurately. Machine learning algorithms were applied each at a time to both the standard SMOTE and enhanced SMOTE-based model. Evaluations of their performance before and after enhancements were conducted.

4.6.1 Comparison of Logistic Regression Algorithm Incorporating SMOTE

Logistic Regression (LR) is used in machine learning for binary (two) classification, predicting the probability of a (two) binary class outcome. The algorithm was applied in credit modelling to classify whether a borrower will default or not based on different features as discussed in previous sections. The results for standard SMOTE were compared with that of enhanced SMOTE-based model. Table 8 below shows how the two models compare in terms of performance metrics.

Table 8

Comparison of Logistic Regression Algorithm Metrics with standard SMOTE model and enhanced SMOTE-based model

Model Performance Metric	Logistic Regression (STANDARD SMOTE)	Logistic Regression (ENHANCED SMOTE)
Accuracy Score	0.6581	0.5311
Kappa Statistic	-0.0002	0.0482
Time Taken (s)	0.0309	414.9764
Mean Absolute Error	0.3419	0.4689
Root Mean Squared Error	0.5847	0.6848
Relative Absolute Error (%)	0.1029	0.1411
Root Relative Squared Error (%)	4.5358	5.3120

The results show that incorporating enhanced SMOTE-based model improved class balance, as seen in the Kappa Statistic, but it led to a decreased accuracy (53%), and an increase in the error rate (47%).

In addition, the Kappa statistic, which shows the agreement between predicted and actual classifications was -0.0002 for Logistic Regression with standard SMOTE, indicating less agreement, while it was 0.0482, indicating an improvement in agreement after balancing the classes in enhanced SMOTE-based model. The time taken in LR with standard SMOTE was higher (414 seconds), showing the increased computational complexity required to generate synthetic samples. The mean absolute error indicated the incorrectly classified values; 34% in LR standard SMOTE and 47% in LR with enhanced SMOTE-based model.

The Root Mean Squared Error (RMSE), which provides the square root of the average squared differences between predicted and the actual values. The RMSE increased from 0.5847 to 0.6848 with enhanced SMOTE-based model, showing a similar trend of

increased error. Finally, the Root Relative Squared Error, which compares the RMSE to the standard deviation of actual values, rose from 4.5358% to 5.3120% with enhanced SMOTE-based model, indicating a higher deviation of predictions from actual values after applying enhancements to standard SMOTE.

4.6.1.1 Comparison of Accuracy, Sensitivity & Specificity of Logistic Regression with Standard SMOTE Model and Enhanced SMOTE-based Model

Further evaluations on accuracy, sensitivity and specificity between standard SMOTE and Enhanced SMOTE-based Model with logistic regression were conducted. Table 9 below shows how these two models compared based on the outlined metrics.

Table 9

Comparison of Accuracy, Sensitivity & Specificity of SMOTE with Logistic Regression

Model Performance Metric	Logistic Regression (standard SMOTE)	Logistic Regression (enhanced SMOTE)
True Positive	486	264
False Negative	2	224
False Positive	251	123
True Negative	1	129
Sensitivity (TP / (TP+FN))	0.9959	0.5410
Specificity = TN / (TN+FP)	0.0040	0.5119
Accuracy (TP+TN)/ (TP+FN+FP+TN)	0.6581	0.5311

The accuracy of the models as seen in was 66% for Logistic Regression with standard SMOTE model and decreased to 53% with enhanced SMOTE-based model. The true positive count, representing correctly predicted defaults, was 486 with standard SMOTE model and 264 with enhanced SMOTE-based model, while false negatives (missed defaults) were 2 with standard SMOTE model and increased to 224 with enhanced

SMOTE-based model. The false positives (incorrectly predicted defaults) were 251 with standard SMOTE model and decreased to 123 with enhanced SMOTE-based model, and true negatives (correctly predicted non-defaults) were 1 with standard SMOTE model, increasing to 129 with enhanced SMOTE-based model. Sensitivity, which measures the ability to correctly identify defaults, was high at 0.9959 with standard SMOTE model but dropped to 0.5410 with enhanced SMOTE-based model. Specificity, indicating the ability to correctly identify non-defaults, was very low at 0.0040 with standard SMOTE model but improved to 0.5119 with enhanced SMOTE-based model. The decrease in accuracy and sensitivity with enhanced SMOTE-based model, accompanied by an increase in specificity, highlights the trade-offs involved in addressing class imbalance. Enhanced SMOTE-based model improved the balance of predictions, particularly for non-defaults, but at the cost of reduced overall predictive accuracy.

4.6.2 Comparison of Decision Trees Algorithm Incorporating SMOTE

Overall, Decision Trees with enhanced SMOTE-based model demonstrated improved accuracy and reduced error rates, indicating better predictive performance in credit modelling. Decision Trees are used for binary classification by providing a visual representation of decisions and their associated consequences. The algorithm was applied in credit modelling to classify whether a borrower will default based on various predictive features. Table 10 below shows how these two models compared based on the outlined metrics.

Table 10

Comparison of Decision Trees Model Metrics with standard SMOTE model and enhanced SMOTE-based model

Model Performance Metric	Decision (standard SMOTE)	Trees (enhanced SMOTE)
Accuracy Score	0.5432	0.8757
Kappa Statistic	-0.0268	0.7237
Time Taken (s)	1274.4364	1283.3060
Mean Absolute Error	0.4568	0.1243
Root Mean Squared Error	0.6758	0.3526
Relative Absolute Error (%)	0.1374	0.0374
Root Relative Squared Error (%)	5.2426	2.7352

The results show that incorporating enhanced SMOTE-based model significantly improved class balance, reflected in the Kappa statistic, which increased from -0.0268 with standard SMOTE model to 0.7237 with enhanced SMOTE-based model. This indicates a substantial improvement in agreement between predicted and actual classifications.

The accuracy score, measuring the proportion of correct predictions, improved from 54.32% with standard SMOTE model to 87.57% with enhanced SMOTE-based model, showing better overall performance with enhanced SMOTE-based model. The mean absolute error (MAE), decreased from 45.68% to 12.43% with enhanced SMOTE-based model, showing fewer errors in classification. The root mean squared error (RMSE), decreased from 0.6758 to 0.3526 with enhanced SMOTE-based model, indicating a reduction in prediction error. The relative absolute error (RAE) percentage also decreased from 13.74% to 3.74%, highlighting a significant reduction in the relative error.

Similarly, the root relative squared error (RRSE), decreased from 5.2426% to 2.7352% with enhanced SMOTE-based model, indicating better model performance.

However, the time taken increased slightly from 1274.4364 seconds to 1283.3060 seconds, reflecting the additional computational effort required for generating synthetic samples.

4.6.2.1 Comparison of Accuracy, Sensitivity & Specificity of Decision Trees with Standard SMOTE Model and Enhanced SMOTE-based Model

Further evaluations on accuracy, sensitivity and specificity between standard SMOTE and enhanced SMOTE-based Model with Decision Trees were conducted. Table 11 below shows how these two models compared based on the outlined metrics.

Table 11

Comparison of Accuracy, Sensitivity & Specificity of SMOTE with Decision Trees

Model Performance Metric	Decision Trees (standard SMOTE)	Decision Trees (Enhanced SMOTE)
True Positive	324	441
False Negative	164	47
False Positive	174	45
True Negative	78	207
Sensitivity (TP / (TP+FN))	0.6639	0.9037
Specificity = TN / (TN+FP)	0.3095	0.8214
Accuracy (TP+TN)/ (TP+FN+FP+TN)	0.5432	0.8757

The results show there is an overall increase in accuracy, specificity, sensitivity while fitting decision trees with enhanced SMOTE-based model. The accuracy of Decision Trees was 54.32% with standard SMOTE model and notably increased to 87.57% with enhanced SMOTE-based model, an increase 34%. This improvement underscores enhanced SMOTE's effectiveness in balancing class distributions, resulting in more accurate predictions overall. Sensitivity increased substantially to 90.37% with enhanced SMOTE-based model as opposed to a sensitivity of 66.39% with standard SMOTE model in Decision Trees. This enhancement indicates that enhanced SMOTE effectively reduced false negatives (missed defaults), thereby improving the model's ability to detect actual defaults. The specificity of the model improved to 82.14% with enhanced SMOTE-based model, improving the model's capability to effectively identify non-defaults from a specificity of 30.95% while fitting Decision Trees with standard SMOTE model. However, it's imperative to note that while enhanced SMOTE improves these metrics related to class imbalance, it can lead to trade-offs such as increased computational complexity.

4.6.3 Comparison of Random Forests Algorithm Incorporating SMOTE

Random forests exhibited the best performance results among all the machine learning algorithms applied in the research. Table 12 shows how well enhanced SMOTE-based model with Random forests algorithms improved on performance metrics.

Table 12

Comparison of Random Forests Model Metrics with standard SMOTE and enhanced SMOTE-based model

Model Performance Metric	Random (standard SMOTE)	Forests (Enhanced SMOTE)
Accuracy Score	0.5919	0.8770
Kappa Statistic	0.0055	0.7249
Time Taken (s)	3760.3831	3814.9702
Mean Absolute Error	0.4081	0.1230
Root Mean Squared Error	0.6388	0.3507
Relative Absolute Error (%)	0.1228	0.0370
Root Relative Squared Error (%)	4.9556	2.7203

The results show that Random forests accuracy increased from 59.19% with standard SMOTE model to 87.70% with enhanced SMOTE-based model in indicating superior overall performance with enhanced SMOTE-based model's class balance enhancement. Random Forests are an ensemble learning method used during ML for classification tasks, combining many decision trees to increase accuracy and minimize overfitting. The model was fitted to the credit data to predict whether a borrower will default based on the predictive features as discussed in previous sections.

The Kappa statistic also improved from 0.0055 with standard SMOTE model to 0.7249 with enhanced SMOTE-based model demonstrating enhanced agreement in classifying defaults and non-defaults with enhanced SMOTE-based model. Random Forests achieved a lower error rate of 12.30% with enhanced SMOTE-based model compared to 40.81% with standard SMOTE model, indicating fewer errors in prediction. Similarly, root mean squared error (RMSE) decreased from 0.6388 to 0.3507 enhanced SMOTE-based model,

indicating improved precision in predicting credit defaults. RAE decreased significantly from 12.28% to 3.70% enhanced SMOTE-based model, RRSE decreased from 4.9556% to 2.7203% with enhanced SMOTE-based model, indicating better overall model performance and precision in predicting credit defaults. The time taken for Random Forests increased slightly from 3760.3831 seconds to 3814.9702 seconds with enhanced SMOTE-based model.

4.6.3.1 Comparison of Accuracy, Sensitivity & Specificity of Random Forests with Standard SMOTE Model and Enhanced SMOTE-based Model

Further evaluations on accuracy, sensitivity and specificity between standard SMOTE and Enhanced SMOTE-based Model with Random forests were conducted. Table 13 below shows how these two models compared based on the outlined metrics.

Table 13

Comparison of Accuracy, Sensitivity & Specificity of Random Forests with standard SMOTE and with enhanced SMOTE-based model

Model Performance Metric	Random Forests (standard SMOTE)	Random Forests (Enhanced SMOTE)
True Positive	382	445
False Negative	106	43
False Positive	196	48
True Negative	56	204
Sensitivity (TP / (TP+FN))	0.7828	0.9119
Specificity = TN / (TN+FP)	0.2222	0.8095
Accuracy(TP+TN)/ (TP+FN+FP+TN)	0.5919	0.8770

The results show there is an overall increase in accuracy, specificity, sensitivity while fitting random forests with enhanced SMOTE-based model. The accuracy of Random Forests increased from 59.19% with standard SMOTE model to 87.70% with enhanced SMOTE-based model, an increase 28%. This improvement underscores enhanced SMOTE's effectiveness in balancing class distributions, resulting in more accurate predictions overall. Sensitivity improved significantly from 78.28% with standard SMOTE model to 91.19% with enhanced SMOTE-based model. This enhancement indicates that enhanced SMOTE effectively reduced false negatives (missed defaults), thereby increasing the model's ability to identify actual defaults. The specificity of the model improved to 80.95% from 22.22% after fitting with enhanced SMOTE-based model, improving the ability of the model in identifying non-defaults. This improvement came with increased time in computation, as seen in the higher amount of time take while fitting Random Forests with enhanced SMOTE-based model.

4.6.3.2 Importance of Predictors Used in the Random Forests Model

This research evaluated further the importance of variables used in the Random Forests with SMOTE. Table 14 below shows the importance of selected variables.

Table 14*Importance of Predictors Used in the Random Forests Model*

Predictor variables (features)	Importance
Loan amount	0.442
Age	0.370
Quarter_year_Q4	0.018
Quarter_year_Q1	0.016
Quarter_year_Q2	0.016
Loan usage_business	0.015
Disburs_method_mobile banking	0.013
Gender_female	0.013
Gender_male	0.013
Disburs_method_manual	0.012
Quarter_year_q3	0.012
Loan usage_others	0.011
Collateral_logbook	0.010
Collateral_checkoff	0.010
Loan usage_education	0.010
Loan usage_agriculture	0.008
Loan usage_medical	0.005
Loan usage_asset purchase	0.004
Loan usage_house development	0.002

Table 14 presents the variable importance in a Random Forest model used for credit modelling. Variable importance is assessed based on the weight of each features

contribution to the model's predictive accuracy. Among the features analyzed, "Loan Amount" and "Age" emerge as the most influential predictors, with importance scores of 0.442 and 0.370, respectively. These variables reflect significant factors in determining credit risk and borrower behaviour. Other variables, such as "quarter year" indicating the quarter of the year, and "LOAN USAGE" categories like "BUSINESS" and "EDUCATION," also show varying degrees of importance, contributing to the model's overall predictive power.

4.6.4 Comparison of Support Vector Machines Algorithm Incorporating SMOTE

SVM is an ML algorithm used for binary classification (or more) by finding the optimal hyperplane in the transformed dimension that separates the data into classes. In this study, SVM was used to predict borrower defaults based on various features. Table 15 shows how standard SMOTE model and enhanced SMOTE-based model compared based on the outlined metrics.

Table 15

Comparison of SVM Model Metrics with standard SMOTE model and enhanced SMOTE-based model

Model Performance Metric	Support Vector Machines (standard SMOTE)	Support Vector Machines (Enhanced SMOTE)
Accuracy Score (%)	0.6635	0.5568
Kappa Statistic (%)	0.0206	0.1196
Time Taken (s)	106293.8592	106337.8267
Mean Absolute Error (%)	0.3365	0.4432
Root Mean Squared Error (%)	0.5801	0.6658
Relative Absolute Error (%)	0.1012	0.1334
Root Relative Squared Error (%)	4.4998	5.1645

The results in Table 15 indicate that the accuracy of Support Vector Machines (SVM) decreased from 66.35% with standard SMOTE model to 55.68% with enhanced SMOTE-based model, reflecting the impact of balancing class distributions. The Kappa statistic improved from 0.0206 with standard SMOTE model to 0.1196 with enhanced SMOTE-based model, indicating better agreement between the predicted and actual classifications when enhanced SMOTE was applied. The mean absolute error (MAE) increased from 33.65% to 44.32% enhanced SMOTE-based model, suggesting a higher error rate in predictions. Similarly, the root mean squared error (RMSE) increased from 0.5801 to 0.6658, indicating less precision in predictions with enhanced SMOTE-based model.

The relative absolute error (RAE) increased from 10.12% to 13.34% with enhanced SMOTE-based model, and the root relative squared error (RRSE) increased from 4.4998% to 5.1645%, suggesting a decrease in overall model precision with the use of enhanced SMOTE-based model. The time taken for the standard SMOTE model slightly increased from 106293.8592 seconds to 106337.8267 seconds, reflecting the computational cost of implementing enhanced SMOTE-based model.

4.6.4.1 Comparison of Accuracy, Sensitivity & Specificity of SVM with Standard SMOTE Model and Enhanced SMOTE-based Model

Further evaluations on accuracy, sensitivity and specificity between standard SMOTE and Enhanced SMOTE-based Model with SVM were conducted. Table 16 below shows how these two models compared based on the outlined metrics.

Table 16

Comparison of Accuracy, Sensitivity & Specificity of SVM with standard SMOTE and enhanced SMOTE-based model

Model Performance Metric	Support Vector Machines (standard SMOTE)	Support Vector Machines (Enhanced SMOTE)
True Positive	486	261
False Negative	2	227
False Positive	247	101
True Negative	5	151
Sensitivity (TP / (TP+FN))	0.9959	0.5348
Specificity = TN / (TN+FP)	0.0198	0.5992
Accuracy(TP+TN)/ (TP+FN+FP+TN)	0.6635	0.5568

In Table 16, the results show a change in accuracy, sensitivity, and specificity for SVM with enhanced SMOTE-based model. The accuracy of SVM decreased from 66.35% with standard SMOTE model to 55.68% with enhanced SMOTE-based model. This decrease highlights the challenges of balancing class distributions while maintaining overall prediction accuracy. Sensitivity, used for identifying true positives (defaults), dropped from 99.59% to 53.48% with enhanced SMOTE-based model, indicating an increase in false negatives.

Specificity, representing the ability to correctly identify true negatives (non-defaults), improved from 1.98% with standard SMOTE model to 59.92% with enhanced SMOTE-based model, demonstrating that enhanced SMOTE-based model effectively increased the ability of the model to identify non-defaults. Although the specificity improved, the

decrease in sensitivity and overall accuracy suggests trade-offs when applying SMOTE to SVM models. The results underscore the complexities of balancing predictive performance across different metrics when addressing class imbalance.

4.6.5 Comparison of Artificial Neural Networks Algorithm Incorporating SMOTE

The model performance of Artificial Neural Networks with standard SMOTE model and enhanced SMOTE-based model was evaluated. Table 17 below shows how the two models compared based on the outlined metrics.

Table 17

Comparison of ANN Model Metrics with standard SMOTE model and enhanced SMOTE-based model

Model Performance Metric	Artificial Network (standard SMOTE)	Neural (standard SMOTE)	Artificial Neural Network (Enhanced SMOTE)
Accuracy Score	0.5595		0.8486
Kappa Statistic	0.0039		0.6650
Time Taken (s)	107308.4930		107583.2157
Mean Absolute Error	0.4405		0.1514
Root Mean Squared Error	0.6637		0.3890
Relative Absolute Error (%)	0.1325		0.0455
Root Relative Squared Error (%)	5.1487		3.0179

The results indicate that the accuracy of the Artificial Neural Network (ANN) increased from 55.95% with standard SMOTE model to 84.86% with enhanced SMOTE-based model, showing the effectiveness of SMOTE in enhancing class balance and overall

model performance. The ANN used the MLPClassifier with 3 hidden layers which allowed the model to learn the complex patterns present in the data.

The Kappa statistic improved from 0.0039 with standard SMOTE model to 0.6650 with enhanced SMOTE-based model, indicating better agreement in classifying defaults and non-defaults after balancing the data. The mean absolute error (MAE) decreased significantly from 44.05% to 15.14% with enhanced SMOTE-based model, reflecting fewer prediction errors. Similarly, the root mean squared error (RMSE) dropped from 0.6637 to 0.3890, demonstrating improved precision in predictions.

Relative absolute error (RAE) decreased from 13.25% to 4.55%, and root relative squared error (RRSE) decreased from 5.1487% to 3.0179% with enhanced SMOTE-based model, indicating enhanced model accuracy and precision. However, the time taken for training the ANN increased slightly from 107308.4930 seconds to 107583.2157 seconds, showing the additional computational effort required when using enhanced SMOTE-based model.

4.6.5.1 Comparison of Accuracy, Sensitivity & Specificity of ANN with Standard SMOTE Model and Enhanced SMOTE-based Model

Further evaluations on accuracy, sensitivity and specificity between standard SMOTE and Enhanced SMOTE-based Model with ANN were conducted. Table 18 below shows how these two models compared based on the outlined metrics.

Table 18

Comparison of Accuracy, Sensitivity & Specificity of ANN with Standard SMOTE Model and Enhanced SMOTE-based Model

Model Performance Metric	Artificial Neural Network (Standard SMOTE)	Artificial Neural Network (Enhanced SMOTE)
True Positive	333	429
False Negative	155	59
False Positive	171	53
True Negative	81	199
Sensitivity (TP / (TP+FN))	0.6824	0.8791
Specificity = TN / (TN+FP)	0.3214	0.7897
Accuracy (TP+TN) / (TP+FN+FP+TN)	0.5595	0.8486

In Table 18, the results show an increase in accuracy, sensitivity, and specificity for the Artificial Neural Network (ANN) with enhanced SMOTE-based model. The accuracy improved from 55.95% with standard SMOTE model to 84.86% with enhanced SMOTE-based model, indicating that enhanced SMOTE-based model effectively balanced class distributions, resulting in more accurate predictions. Sensitivity increased from 68.24% to 87.91% with enhanced SMOTE-based model, showing a significant reduction in false negatives.

Specificity improved from 32.14% with standard SMOTE model to 78.97% with enhanced SMOTE-based model, reflecting better identification of non-defaults. The improvement in these metrics highlights the effectiveness of enhanced SMOTE-based model in enhancing the ANN model's predictive performance, although at the cost of increased computation.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

This chapter provides the summary of the research, the conclusions, the recommendations and the suggestions for further research.

5.1 Summary of Key Findings

The following objectives guided the investigation in this study; first the study aimed to investigate through a survey the existing credit Modelling techniques by financial institutions in Meru County. The second objective sought to develop and propose enhancements to the SMOTE-based model while the third objective was to evaluate and validate the effectiveness enhanced SMOTE-based model in enhancing credit Modelling accuracy using data from the lab.

The first goal was to conduct a survey of the existing credit modelling techniques. The field research into credit modelling techniques in Meru County's financial institutions revealed a reliance on traditional credit scoring models and reference bureau data. While some banks employ advanced quantitative models like default risk and non-performing loan portfolios, many face challenges due to the complexity and technical demands of these models, requiring extensive staff training. This indicates a varied adoption of credit-risk management practices across different banks.

The second goal of this research was to develop an enhanced SMOTE-based model for enhancing the predictive accuracy of credit modelling. The model aimed at addressing class imbalance in the training set and addressing overfitting challenge that affect model's predictive accuracy in credit modelling. The enhanced model demonstrated improved class balance and addressed the challenge of overfitting that resulted to its overall improvement of predictive accuracy. Applying the enhanced model to logistic regression showed improved class balance but decreased accuracy and increased error rates, while decision trees and random forests demonstrated substantial improvements in accuracy and reduced errors. Support vector machines saw a decrease in accuracy but an

improvement in specificity, and artificial neural networks showed significant gains in accuracy, sensitivity, and specificity, highlighting the varied impact of SMOTE across different models.

The third goal aimed to evaluate the overall effectiveness of the enhanced SMOTE-based model in enhancing accuracy in credit modelling from a comparative analysis of various machine learning algorithms using standard SMOTE model and an enhanced SMOTE-based Model. Key findings from this evaluation revealed the following:

1. Logistic Regression

The incorporation of enhanced SMOTE-based model with Logistic Regression showed a mixed impact. While SMOTE improved the class balance, as evidenced by an increase in the Kappa statistic from -0.0002 to 0.0482, it resulted in a significant decrease in accuracy from 65.81% to 53.11%. Additionally, error rates increased, with the mean absolute error rising from 34.19% to 46.89%. The increased computational complexity, reflected in the significant rise in time taken (from 0.0309 seconds to 414.9764 seconds), further highlights the trade-offs involved. These results suggest that while SMOTE can enhance agreement between predicted and actual classifications, its impact on overall accuracy and error rates in Logistic Regression is detrimental, potentially due to the nature of the algorithm and its sensitivity to data balance adjustments. Studies by Fernández et al. (2018) and López et al. (2013) have shown similar mixed results, emphasizing the variability in Logistic Regression's response to SMOTE.

2. Decision Trees

Decision Trees exhibited substantial improvements with the application of enhanced SMOTE-based model. Accuracy increased from 54.32% to 87.57%, and the Kappa statistic rose from -0.0268 to 0.7237, indicating a much better agreement between predicted and actual classifications. Error rates decreased significantly, with the mean absolute error dropping from 45.68% to 12.43%, and the root mean squared error reducing from 0.6758 to 0.3526. Sensitivity and specificity also improved dramatically, showcasing the model's enhanced ability to correctly identify both defaults and non-defaults. These findings highlight Decision Trees' robustness in handling class

imbalances with SMOTE, leading to improved predictive performance and reliability. Similar positive outcomes were reported by Chawla et al. (2002) and Bellinger et al. (2020), who found that Decision Trees benefit significantly from SMOTE in terms of both accuracy and error rates.

3. Random Forests

Similar to Decision Trees, Random Forests demonstrated marked improvements with enhanced SMOTE-based model. Accuracy increased from 59.19% to 87.70%, and the Kappa statistic improved from 0.0055 to 0.7249, indicating better classification agreement. Error rates showed significant reductions, with the mean absolute error decreasing from 40.81% to 12.30%, and the root mean squared error dropping from 0.6388 to 0.3507. The enhancements in sensitivity (from 78.28% to 91.19%) and specificity (from 22.22% to 80.95%) further underscore the model's effectiveness in handling imbalanced datasets with SMOTE. These results suggest that Random Forests, when combined with enhanced SMOTE-based model, can significantly improve the accuracy and precision of credit risk predictions. These findings align with those of Wang et al. (2017) and Ganganwar (2012), who observed similar improvements in Random Forest performance with SMOTE.

4. Support Vector Machines (SVM)

The impact of enhanced SMOTE-based model on SVM was less favourable. While the Kappa statistic improved from 0.0206 to 0.1196, indicating better classification agreement, accuracy decreased from 66.35% to 55.68%. Error rates also increased, with the mean absolute error rising from 33.65% to 44.32%, and the root mean squared error increasing from 0.5801 to 0.6658. Despite an improvement in specificity from 1.98% to 59.92%, the decrease in sensitivity from 99.59% to 53.48% highlights the challenges of balancing class distributions while maintaining overall prediction accuracy. These findings suggest that SVM may not be as well-suited for handling imbalanced datasets with SMOTE compared to other models. Research by Tang et al. (2009) and Barandela et al. (2003) supports this conclusion, showing that SVM can be sensitive to the balance of the dataset and might not always benefit from oversampling techniques like SMOTE.

5. Artificial Neural Networks (ANN)

ANN showed significant benefits when applied to enhanced SMOTE-based model. Accuracy improved from 55.95% to 84.86%, and the Kappa statistic increased from 0.0039 to 0.6650, indicating better classification agreement. Error rates decreased markedly, with the mean absolute error dropping from 44.05% to 15.14%, and the root mean squared error reducing from 0.6637 to 0.3890. Sensitivity and specificity both improved, highlighting the model's enhanced ability to correctly identify defaults and non-defaults. These findings demonstrate the effectiveness of enhanced SMOTE-based model in improving ANN's predictive performance, making it a robust choice for credit risk modelling with imbalanced data. Similar results were found by Liu et al. (2008) and Fernández et al. (2018), who noted significant improvements in ANN performance with SMOTE application.

Generally, Random Forests with enhanced SMOTE-based model demonstrated the most significant improvement in accuracy, sensitivity, and specificity. This model effectively addressed class imbalance, resulting in enhanced predictive performance for credit default prediction.

5.2 Conclusions

This study explored credit Modelling techniques in Meru County, highlighting the prevalence of traditional credit scoring models and the varying adoption of quantitative approaches among financial institutions. It underscored the challenges posed by these models, including complexity and the need for extensive training. The application of enhanced SMOTE to address class imbalance and address overfitting showed promising results, particularly in Decision Trees and Random Forests, which demonstrated enhanced predictive accuracy and balance.

Decision Trees and Random Forests emerged as effective models when integrated with the enhanced SMOTE-based model, showcasing improved predictive performance in credit default prediction. Decision Trees, for instance, significantly increased accuracy, while Random Forests excelled in both sensitivity and specificity. These findings suggest that leveraging enhancements to SMOTE in these models can mitigate the challenges of class imbalance and overfitting, providing robust tools for credit risk management in Meru County.

In conclusion, the study advocates for the adoption of SMOTE-enhanced Random Forests in credit Modelling due to their balanced performance and enhanced predictive capabilities. Despite computational complexities, these models offer reliable solutions for financial institutions seeking to optimize credit risk assessment strategies.

5.3 Recommendations

To achieve high accuracy in credit modelling, it is advisable to use Decision Trees, Random Forests, or ANN in enhanced SMOTE-based models to leverage their improved performance in handling imbalanced datasets. These models demonstrated significant benefits from SMOTE, enhancing their predictive accuracy and reliability. It is essential to prepare for the increased computational resources required for training these models with SMOTE, as they can be more resource-intensive.

When applying SMOTE to Logistic Regression or SVM, caution is recommended due to potential reductions in accuracy and increases in computational complexity. These models may be more appropriate in scenarios where class balance is less critical or where alternative techniques for handling imbalance are employed.

Further research into advanced techniques for class imbalance, such as hybrid approaches or other resampling methods, could provide additional improvements in predictive performance. Additionally, exploring the effects of feature selection and dimensionality reduction on model performance with SMOTE could lead to refined and more accurate results.

Regular evaluation of models using a comprehensive set of performance metrics is crucial to understand the trade-offs involved and to select the most appropriate model for specific credit risk assessment needs. Notably, balancing class distributions effectively while maintaining high accuracy and precision remain a key consideration in the application of SMOTE and other data balancing techniques.

REFERENCES

- Abedin, M., Guotai, C., & Hajek, P. E. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex Intelligent Systems*. <https://doi.org/10.1007/s40747-022-00421-8>
- Ahmad, H., Kasasbeh, B., Al-Dabaybah, B., & Faisal, E. (2023). An effective oversampling technique for credit card fraud detection by utilizing noise filtering and fuzzy c-means clustering. *International Journal of Data and Network Science*, 7(2), 1-8. <https://doi.org/10.5267/j.ijdns.2023.1.002>
- Akkoc, S. (2012). An empirical comparison of conventional techniques, neural networks, and the three-stage hybrid adaptive neuro fuzzy inference system (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, 218(1), 168-178. <https://doi.org/10.1016/j.ejor.2011.10.003>
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589-609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- Alvarez, J., & Bansal, A. (2021). Comparative analysis of SMOTE variants for financial default prediction. *Applied Soft Computing*, 102, 106889. <https://doi.org/10.1016/j.asoc.2021.106889>
- Anderson, R. (2007). *The credit scoring toolkit*. Oxford University Press.
- Banerjee, S., & Dutta, P. (2024). Comparative analysis of SMOTE variants in credit risk assessment. *Journal of Risk Finance*, 25(2), 134-150. <https://doi.org/10.1108/JRF-01-2024-0013>
- Bauer, K., & Tharakan, J. (2006). Evaluating the performance of credit scoring models: A review. *Journal of Banking & Finance*, 30(6), 1825-1836. <https://doi.org/10.1016/j.jbankfin.2005.10.003>

- Bekhet, H., & Eletter, S. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, 4(1), 20-28. <https://doi.org/10.1016/j.rdf.2014.02.002>
- Bellinger, C., Reinke, C., & Ahmed, M. (2020). Improving the accuracy of credit risk modeling using boosting techniques. *Expert Systems with Applications*, 145, 113104. <https://doi.org/10.1016/j.eswa.2020.113104>
- Caruso, G., Gattone, S., Fortuna, F., & Di Battista, T. (2021). Cluster analysis for mixed data: An application to credit risk evaluation. *Socio-Economic Planning Sciences*, 73, 100994. <https://doi.org/10.1016/j.seps.2021.100994>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. <https://doi.org/10.1613/jair.953>
- Chopra, A., & Bhilare, P. (2018). Application of ensemble models in credit scoring models. *Business Perspectives and Research*, 6(2), 129-141. <https://doi.org/10.1177/2278533718775578>
- Cisko, Š., & Klieštik, T. (2013). *Finančný manažment podniku II*. Zilina: EDIS Publishers, University of Žilina.
- Correa, B. A. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, 64, 134-132. <https://doi.org/10.1016/j.eswa.2016.01.016>
- Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447-1466. <https://doi.org/10.1016/j.ejor.2005.07.010>
- Elyan, E., Moreno-Garcia, C., & Jayne, C. (2020). CDSMOTE: Class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. *Neural Computing and Applications*, 32(12), 2839-2851. <https://doi.org/10.1007/s00500-019-03932-0>

- Feng, S., Xingchao, Z., Gang, K., & Fawaz, E. A. (2021). A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Applied Soft Computing*, 106, 107281. <https://doi.org/10.1016/j.asoc.2021.107281>
- Fernandes, A., & Freitas, G. (2022). Heart disease prediction and classification using machine learning. *ResearchGate*. <https://doi.org/10.13140/RG.2.2.14778.04805>
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer.
- Guidolin, M., & Pedio, M. (2021). Sharpening the accuracy of credit scoring models with machine learning algorithms. In *Data Science and Economic Finance: Methodology and Applications* (pp. 89-115). Springer. https://doi.org/10.1007/978-3-030-66891-4_5
- Gul, F., & Kahn, M. (2018). The role of expert judgment in credit risk assessment: A case study. *Journal of Banking & Finance*, 88, 1-10. <https://doi.org/10.1016/j.jbankfin.2017.11.002>
- Guo, H., Li, Y., Shang, J., Gu, M., Huang, Y., & Gong, B. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 83, 220-239. <https://doi.org/10.1016/j.eswa.2016.09.026>
- Hamal, S., & Senvar, O. (2021). Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs. *International Journal of Computational Intelligence Systems*, 14(1), 769–782. <https://doi.org/10.2991/ijcis.d.210203.007>
- Heiat, A. (2012). Comparing performance of data mining models for computer credit scoring. *Journal of International Finance and Economics*, 12(1), 78-83.
- Huang, Z., & Chen, H. (2020). Credit risk modeling: A comprehensive review. *Journal of Risk Finance*, 21(1), 65-86. <https://doi.org/10.1108/JRF-05-2019-0097>

- Jagelid, M., & Movin, M. (2021). A comparison of resampling techniques to handle the class imbalance problem in machine learning: Conversion prediction of Spotify users—A case study. *Online*. <https://doi.org/10.13140/RG.2.2.23779.11041>
- Javad, H. J., Abdolreza, M., Mohammad, A. N., Solomon, S. O., & Sadiq, H. (2023). Effective class-imbalance learning based on SMOTE and convolutional neural networks. *Journal of Applied Sciences*, *13*(3), 4006. <https://doi.org/10.3390/app13034006>
- Johnson, J., & Khoshgoftaar, T. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1-54. <https://doi.org/10.1186/s40537-019-0192-5>
- Kaur, P., & Gosain, A. (2018). Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. In *Advances in Intelligent Systems and Computing* (pp. 23-30). Springer. https://doi.org/10.1007/978-3-030-19192-7_3
- Khan, M. A., & Hussain, S. (2022). Application of SMOTE variants in banking credit risk models. *Finance Research Letters*, *45*, 102247. <https://doi.org/10.1016/j.frl.2022.102247>
- Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, *83*, 105645. <https://doi.org/10.1016/j.asoc.2019.105645>
- Kulkarni, S. V., & Dhage, S. N. (2019). Advanced credit score calculation using social media and machine learning. In *Soft Computing and Intelligent Systems: Techniques and Applications* (pp. 2373–2380). https://doi.org/10.1007/978-3-030-22514-7_22
- Li, W., Ding, S., Chen, Y., Wang, H., & Yang, S. (2019a). Transfer learning-based default prediction model for consumer credit in China. *The Journal of Supercomputing*, *75*(2), 862-884. <https://doi.org/10.1007/s11227-018-2542-3>

- Ling, C. X., & Sheng, V. (2015). Cost-sensitive learning and the class imbalance problem. In *Proceedings of the 2015 IEEE International Conference on Data Mining* (pp. 225-230). <https://doi.org/10.1109/ICDM.2015.100>
- Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2nd ed.). Stata Press.
- Lu, W. (2022). Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization. *Applied Soft Computing*, *121*, 106850. <https://doi.org/10.1016/j.asoc.2022.106850>
- Ma, Y., & He, H. (2013). *Imbalanced learning: Foundations, algorithms, and applications*. Hoboken, NJ: John Wiley & Sons.
- Maha, B., Tony, B., & Niall, A. (2020). Identification of credit risk based on cluster analysis of account behaviors. *Journal of the Operational Research Society*, *71*(5), 775-783. <https://doi.org/10.1080/01605682.2019.1613631>
- Mahmoud, A., & Khaled, A. (2023). Ensemble approaches to credit risk prediction using SMOTE. *Journal of Financial Risk Management*, *12*(3), 245-261. <https://doi.org/10.4236/jfrm.2023.123015>
- Mandala, I., Nawangpalupi, C. A., & Praktikto, F. R. (2012). Assessing credit risk: An application of data mining in a rural bank. *Procedia Economics and Finance*, *4*, 406-412. [https://doi.org/10.1016/S2212-5671\(12\)00290-6](https://doi.org/10.1016/S2212-5671(12)00290-6)
- Masmoudi, A., Masmoudi, M. K., Abid, L., & Masmoudi, A. (2019). Credit risk modeling using Bayesian network with a latent variable. *Expert Systems with Applications*, *113*, 157-166. <https://doi.org/10.1016/j.eswa.2018.06.024>
- McKinsey & Company. (2015). Credit risk modeling: The importance of judgment. Retrieved from <https://www.mckinsey.com/industries/financial-services/our-insights>
- Moula, F., Guotai, C., & Abedin, M. (2017). Credit default prediction modeling: An application of support vector machine. *Risk Management*, *19*(3), 158-187. <https://doi.org/10.1057/s41283-017-0008-6>

- Nguyen, T. H., & Tran, Q. (2023). Predicting loan defaults: A SMOTE approach. *International Journal of Banking, Accounting, and Finance*, 13(1), 54-70. <https://doi.org/10.1504/IJBFA.2023.128305>
- Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109-131. <https://doi.org/10.2307/2490395>
- Patel, H., Singh Rajput, D., Thippa Reddy, G., Iwendi, C., Kashif Bashir, A., & Jo, O. (2020). A review on classification of imbalanced data for wireless sensor networks. *International Journal of Distributed Sensor Networks*, 16(10), 1-20. <https://doi.org/10.1177/1550147720966648>
- Patel, R., & Sharma, V. (2024). Real-world applications of SMOTE in banking sector credit risk modeling. *International Journal of Financial Services Management*, 8(4), 302-316. <https://doi.org/10.1504/IJFSM.2024.132001>
- Shashi, D., Handa, S., & Singh, N. (2017). A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. *Expert Systems*, 34(2), e12241. <https://doi.org/10.1111/exsy.12241>
- Sumathi, S., & Sivanandam, S. (2006). *Introduction to data mining and its applications*. Berlin: Springer-Verlag.
- Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, 41(10), 2353–2361. <https://doi.org/10.1016/j.eswa.2013.12.027>
- Yong, S., Huakun, Q., Qianqian, C., Jingming, Z., Jingru, L., Zhengmin, K., & Shuai, W. (2022). Borderline SMOTE algorithm and feature selection-based network anomalies detection strategy. *Energies*, 15(13), 4571. <https://doi.org/10.3390/en15134571>
- Zanin, M. (2016). Combining complex networks and data mining: Why and how. *Physics Reports*, 641, 1-44. <https://doi.org/10.1016/j.physrep.2016.07.001>

Zhang, C., & Wang, Z. (2018). WSMOTE: A novel SMOTE variant for imbalanced data learning. *Knowledge-Based Systems*, 138, 91-103.
<https://doi.org/10.1016/j.knosys.2017.12.036>

Zheng, M., Li, T., Sun, L., Wang, T., Jie, B., Yang, W., & Lv, C. (2021). An automatic sampling ratio detection method based on genetic algorithm for imbalanced data classification. *Knowledge-Based Systems*, 216, 106726.
<https://doi.org/10.1016/j.knosys.2021.106726>

APPENDICES

Appendix 1: Sample Questionnaire

Section 1: Participant Information

1. Please provide your demographic information: (Tick where appropriate)

a. Age:

20-25

25-30

(30-35)

(35-40)

(45-50)

(Above 50)

(I prefer not to answer)

b. Gender: Male Female I prefer not to answer

c. Job Title: Bank credit Data analyst Credit officer

I prefer not to answer

d. Years of Experience in Credit data analysis: (Below 5) (5-10)

(Above 10) I prefer not to answer

Section 2: Familiarity with Credit Modelling Techniques

2. Are you familiar with credit modelling techniques used in banks?

a. Yes b. No

3. If yes, please specify the credit modelling techniques you are familiar with:

.....

Section 3: Understanding of SMOTE and Credit Modelling

4. Have you heard about the Synthetic Minority Oversampling Technique (SMOTE)?
- a. Yes b. No
5. If yes, how would you rate your understanding of SMOTE?
- a. Very familiar
- b. Moderately familiar
- c. Somewhat familiar
- d. Not familiar
6. How important do you think credit modelling is for banks in assessing credit risk?
- a. Very important
- b. Moderately important
- c. Somewhat important
- d. Not important

Section 4: Use of Credit Modelling Techniques in Banks

7. Does your bank currently utilize credit modelling techniques for credit risk assessment?
- a. Yes b. No
8. If yes, which credit modelling techniques are commonly used in your bank?
(Select all that apply)
- a. Logistic Regression
- b. Decision Trees
- c. Random Forest
- d. Support Vector Machines

e. Gradient Boosting f. Other (Please specify):.....

9. How satisfied are you with the accuracy of the current credit modelling techniques in your bank?

a. Very satisfied

b. Moderately satisfied

c. Somewhat satisfied

d. Not satisfied

Section 5: Perceived Benefits and Challenges of SMOTE-based Credit Model

10. Do you believe that implementing an SMOTE-based credit model can improve the accuracy of credit risk assessment in banks?

a. Yes b. No c. Unsure

11. What potential benefits do you foresee in using an SMOTE-based credit model? (Select all that apply)

a. Improved accuracy in predicting credit risk

b. Better identification of potential defaulters

c. Enhanced decision-making for loan approvals

d. Reduction in false positives/negatives

e. Other (Please specify):

12. What challenges or concerns do you anticipate in implementing an SMOTE-based credit model in your bank? (Open-ended)

.....

Section 6: Potential Adoption and Implementation

13. Would your bank be willing to adopt and implement an SMOTE-based credit model?

- a. Yes b. No c. Unsure

14. If yes or unsure, what factors would influence your bank's decision to adopt an SMOTE-based credit model? (Open-ended)

.....

Section 7: Suggestions and Feedback

15. Do you have any suggestions, recommendations, or additional comments regarding the implementation or usage of an SMOTE-based credit model in banks? Please state them below. (Open-ended)

.....

.....

Section 8: Conclusion

16. In your opinion, how likely is the adoption of SMOTE-based credit Modelling techniques in banks in the near future?

- a. Very likely
b. Moderately likely
c. Somewhat likely
d. Not likely

17. Thank you for your participation! If you would like to provide any further comments or suggestions related to this research, please do so below. (Open-ended)

.....

Appendix 2: A sample of Importing Libraries

```
{
  "cells": [
    {
      "cell_type": "code",
      "execution_count": 1,
      "id": "7001b9fd",
      "metadata": {},
      "outputs": [],
      "source": [
        "#import libraries\n",
        "import numpy as np\n",
        "import pandas as pd\n",
        "import matplotlib.pyplot as plt\n",
        "%matplotlib inline\n",
        "from sklearn.linear_model import LogisticRegression as LR\n",
        "from sklearn.model_selection import train_test_split\n",
        "from sklearn.metrics import confusion_matrix\n",
        "from sklearn.metrics import roc_curve"
      ]
    },
    {
      "cell_type": "code",
      "execution_count": 2,
      "id": "555e37f2",
      "metadata": {},
      "outputs": [],
      "source": [
        "data=pd.read_csv('dickson loan.csv')"
      ]
    }
  ],
}
```

Appendix 3: A sample of Creating Random Samples

```

"text/plain": [
    " Collateral Loan
Amount LOAN USAGE Gender Age
Disburs_Method Default \\n",
    "0 Checkoff
10000.0 BUSINESS Male 29
Manual Yes \n",
    "1 Checkoff
10000.0 MEDICAL Male 31
Manual Yes \n",
    "2 Checkoff
10283.0 EDUCATION Male 49
Mobile Banking Yes \n",
    "3 Checkoff
10391.0 EDUCATION Male 27
Mobile Banking Yes \n",
    "4 Checkoff
10427.0 OTHERS Male 36
Mobile Banking Yes \n",
    "\n",
    " quarter_year \n",
    "0 Q2 \n",
    "1 Q3 \n",
    "2 Q4 \n",
    "3 Q4 \n",
    "4 Q2 "
]
},
"execution_count": 3,
"metadata": {},
"output_type":
"execute_result"
}
],
"source": [
"data.head()"
]
},
{
"cell_type": "code",
"execution_count": 4,
"id": "eeb533dd",
"metadata": {},
"outputs": [
{
"data": {
"text/plain": [
(2507, 20)
]
},
"execution_count": 4,
"metadata": {},
"output_type":
"execute_result"
}
],
"source": [
"data.shape"
]
},
{
"cell_type": "code",
"execution_count": 5,
"id": "f355b5f6",
"metadata": {},
"outputs": [
{
"data": {
"text/plain": [
Collateral
object\n",
Loan Amount
float64\n",
LOAN USAGE
object\n",
Gender
object\n",
Age
int64\n",
Disburs_Method
object\n",
Default
object\n",
quarter_year
object\n",
dtype: object"
]
},
"execution_count": 5,
"metadata": {},
"output_type":
"execute_result"
}
]
}

```

Appendix 4: Metrics Calculation


```
"# Metrics calculation\n",
"metrics = {\n",
"    \"Accuracy Score\": accuracy,\n",
"    \"Kappa Statistic\": kappa,\n",
"    \"Time Taken (s)\": time_taken,\n",
"    \"Mean Absolute Error\": mae,\n",
"    \"Root Mean Squared Error\": rmse,\n",
"    \"Relative Absolute Error (%)\": relative_absolute_error,\n",
"    \"Root Relative Squared Error (%)\":
root_relative_squared_error,\n",
"    \"True Positive\": tp,\n",
"    \"False Negative\": fn,\n",
"    \"False Positive\": fp,\n",
"    \"True Negative\": tn,\n",
"    \"Sensitivity\": sensitivity,\n",
"    \"Specificity\": specificity,\n",
"}\n",
"\n",
"# Create a DataFrame\n",
"df_metrics = pd.DataFrame(list(metrics.items()),
columns=[\"Metric\", \"Value\"])\n",
"\n",
"# Export to Word\n",
"doc = Document()\n",
"doc.add_heading(\"Model Performance Metrics\", level=1)\n",
"\n",
"# Add table\n",
"table = doc.add_table(rows=1, cols=2)\n",
"hdr_cells = table.rows[0].cells\n",
"hdr_cells[0].text = \"Metric\"\n",
"hdr_cells[1].text = \"Value\"\n",
"\n",
"for metric, value in metrics.items():\n",
"    row_cells = table.add_row().cells\n",
"    row_cells[0].text = metric\n",
"    row_cells[1].text = f\"{value:.4f}\"\n",
"\n",
"doc.save(\"C:/Users/John/Desktop/kalonje2/logistic2_performance_metrics.docx\")
]
},
{
"cell_type": "code",
"execution_count": 107,
"id": "cc9bccal",
"metadata": {},
"outputs": [],
"source": [
"#fitting logistic with SMOTE #note the accuracy goes down with
smote\n",
"r22=LR().fit(X_res,y_res)\n",
"YP=r22.predict(XTEST)\n",
"\n",
"# Calculate time taken\n",
```

```

"time_taken = time.time() - start_time\n",
"\n",
"# Metrics calculation\n",
"accuracy = accuracy_score(YTEST, YP)\n",
"kappa = cohen_kappa_score(YTEST, YP)\n",
"mae = mean_absolute_error(YTEST, YP)\n",
"rmse = mean_squared_error(YTEST, YP, squared=False)\n",
"relative_absolute_error = mae / sum(abs(YTEST - YTEST.mean())) *
100\n",
"root_relative_squared_error = rmse / (sum((YTEST -
YTEST.mean())**2) ** 0.5) * 100\n",
"\n",
"# Confusion matrix metrics\n",
"tn, fp, fn, tp = confusion_matrix(YTEST, YP).ravel()\n",
"sensitivity = tp / (tp + fn)\n",
"specificity = tn / (tn + fp)"
]
},

```

Appendix 5: Tharaka University Introductory Letter

THARAKA		UNIVERSITY
P.O BOX 193-60215, MARIMANTI, KENYA		Telephone: +(254)-0202008549 Website: https://tharaka.ac.ke Social Media: tharakauni Email: info@tharaka.ac.ke
OFFICE OF THE DIRECTOR BOARD OF POSTGRADUATE STUDIES		

REF: TUN/BPGS/PL/03/24 12th March, 2024

To Whom It May Concern

Dear Sir/Madam,

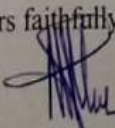
RE: DICKSON MURITHI ADMISSION NUMBER SMT22/03135/21


Mr. Dickson Murithi is a postgraduate student at Tharaka University undertaking a Master degree in **Computer Science**. The student has completed his coursework and is expected to proceed for collection of data after successfully defending his proposal at faculty level. The title of the study is "*Effective Synthetic Minority Oversampling Technique (SMOTE) Based Model for Enhancing Accuracy in Credit Modeling.*" The proposed study will be carried out in **Meru County**.

Any assistance accorded to him will be highly appreciated.

Thank you in advance.

Yours faithfully,


Dr. Marciano Mutiga, Ph.D.
Director, Board of Postgraduate Studies.



Appendix 6: Ethics Review Letter

THARAKA

P.O BOX 193-60215,
MARIMANTI, KENYA



UNIVERSITY

Telephone: +(254)-0202008549
Website: <https://tharaka.ac.ke>
Social Media: tharakauni
Email: info@tharaka.ac.ke

INSTITUTIONAL SCIENTIFIC AND ETHICS REVIEW COMMITTEE

20th February, 2024.

REF: TUNISERC/NSEC/M011

Dear, Dickson Murithi

RE: A smote Based Model for Improving Accuracy in Credit Modelling.

This is to inform you that *Tharaka University ISERC* has reviewed and approved your above research proposal. Your application approval number is *ISERC04023*. The approval period is *20th February 2024 – 20th February, 2025*.

This approval is subject to compliance with the following requirements;

- i. Only approved documents including (informed consents, study instruments, MTA) will be used
- ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by *Tharaka University ISERC*.
- iii. Death and life threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to *Tharaka University ISERC* within 72 hours of notification
- iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to *Tharaka University ISERC* within 72 hours
- v. Clearance for export of biological specimens must be obtained from relevant institutions.
- vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
- vii. Submission of an executive summary report within 90 days upon completion of the study to *Tharaka University ISERC*.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI)

<https://research-portal.nacosti.go.ke> and also obtain other clearances needed.

Yours sincerely,



Dr. Fidelis Ngugi
Chair, ISERC Tharaka University

Appendix 7: NACOSTI License

Republic of Kenya
NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Ref No: 469691

RESEARCH LICENSE




This is to Certify that **Mr. DICKSON MURITHI** of Tharaka University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Meru on the topic: **EFFECTIVE SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) BASED MODEL FOR ENHANCING ACCURACY IN CREDIT MODELING** for the period ending : 17/April/2025.

License No: NACOSTI/P/24/34050

Applicant Identification Number: 469691

Director General
NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Verification QR Code



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

See overleaf for conditions

Appendix 8: Meru County Map

