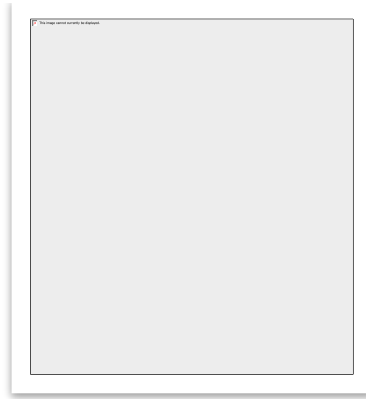


BINDURA UNIVERSITY OF SCIENCE EDUCATION
DEPARTMENT OF MATHEMATICS AND STATISTICS
FACULTY OF SCIENCE AND ENGINEERING



**Modeling The Probability Of Loan Default Via Logistic Regression And Survival Analysis
Techniques**

BY

Rebecca Tafadzwa Mbengi

B202857B

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

SUPERVISOR: MS J. PAGAN'A

JUNE 2024

APPROVAL FORM

This is to certify, that this research project is the result of my own research work and has not been copied or extracted from past sources without acknowledgement. I hereby declare that no part of it has been presented for another degree in this University or elsewhere.



10/06/2024

REBECCA T. MBENGI

.....

.....

B202857B

Signature

Date

Certified by

11/06/2024

Ms.J PAGAN'A

.....

.....

Supervisor

Signature

Date

DR MAGODORA



11/06/2024

Chairperson

Signature

Date

DEDICATION

I dedicate this study to my beloved mother, Sarudzai Machina, for her undying support and unconditional love. This is dedicated to all her hard work and towards the fulfillment of all her unattained dreams if any.

I also dedicate this report to my late father, Donald Mbengi, whose spirit and guidance have been my constant source of strength and inspiration.

To my sister's kids, Sanchez Nigel Chizema and Sancho Nathaniel Chizema, everything is possible through hard work and determination.

ACKNOWLEDGEMENT

With heartfelt gratitude, I acknowledges the invaluable support and guidance that made the completion of this thesis possible.

First and foremost I would like to express my deepest appreciation to my former branch manager, Mrs T Murimigwa for her assistance with this research project data. Her unwavering belief in me and constant encouragement pushed me to pursue this academic journey. Her mentorship and faith in my abilities have been instrumental in shaping me into the scholar I am today.

To my supervisor Ms J. Pagan'a words fall short in conveying my sincere thankfulness. Her patience, expertise, and unwavering dedication to my success have been pillars of strength throughout this process. Her insightful feedback and steadfast support have been truly transformative. I am going to also thank every instructor at the Statistics department for their advice and encouragement along this endeavor.

I am profoundly grateful to the HLF Foundation, whose generosity and vision provided the financial support that made my tertiary education possible. This opportunity has been life-changing, and I is forever indebted to the Foundation for their investment in my future.

To my beloved mother and family, thank you for your unconditional love, prayers, and steadfast support. Your unwavering belief in me has been a constant source of inspiration, and I am honored to make you proud.

I would also like to express my heartfelt appreciation to my best friend, whose companionship, encouragement, and countless hours of moral support have been a beacon of light during the challenges of this endeavor.

I am also grateful to my fellow students, who have become a cherished family. Your camaraderie, shared experiences, and mutual encouragement have enriched this journey immeasurably.

Among all, I thank the Heavenly Father for providing me with the ability to complete my studies.

ABSTRACT

Prediction of loan defaults is an essential component of credit risk assessment, which informs the decision-making processes of financial institutions. This empirical research project aimed to develop a predictive model for estimating the default risk of a loan portfolio by analyzing historical loan data and borrower characteristics. The researcher utilized logistic regression and survival analysis methods to analyze a vast dataset of loan portfolios obtained from KCI Management Consultants. The study's results demonstrated that both logistic regression and survival analysis strategies have relatively similar performance based on the Receiver Operating Characteristic assessment. However, the survival model outperformed the logistic regression method in accurately predicting defaulted and non-defaulted loan portfolios. This suggests that survival analysis strategy offers a viable alternative to the conventional logistic regression method used in assessing credit risk in the MFI sector. The research project also revealed that survival analysis provides several advantages for credit risk management and capital management. By modeling the time to default, survival analysis can help identify key risk indicators before they impact credit markets and provide insights into the relationship between default and borrower characteristics. The study confirmed the importance of utilizing empirical approaches to credit risk management and showcased the advantages of employing survival analysis over logistic regression as a predictive model for loan default risk. These insights provide valuable information for financial institutions looking for a more accurate and effective way of addressing credit risk management.

TABLE OF CONTENTS

APPROVAL FORM	ii
DEDICATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT.....	v
TABLE OF CONTENTS.....	vi
LIST OF FIGURES	ix
LIST OF TABLES	x
LIST OF ACRONYMS	xi
CHAPTER ONE: INTRODUCTION.....	1
1.0 Introduction	1
1.1 Background to the Study	1
1.2 Statement of the Problem.....	3
1.3: Research objectives	4
1.4 Research Questions	4
1.5 Scope of the Study.....	4
1.6 Significance of the study.....	5
1.7 Assumptions of the study	5
1.8 Limitations of the Study.....	5
1.9 Definition of terms	6
1.10 Conclusion.....	7
CHAPTER TWO: LITERATURE REVIEW	8
2.0 Introduction	8
2.1 Theoretical Literature	8
2.2 Credit Risk Management Theory	8
2.3 Information Asymmetry Theory	10
2.4 Empirical Literature	11
2.5 Borrower-Level Factors	12
2.6 Institutional Factors.....	12
2.7 Combining Techniques	12

2.8 Contextual Factors.....	13
2.9 Chapter Conclusion.....	13
CHAPTER THREE: RESEARCH METHODOLOGY	14
3.0: Introduction.....	14
3.1 Research Design.....	14
3.2 Data Sources.....	15
3.3 Target population and sampling methods	15
3.4 Research Instruments	16
3.5 Methods for data collection.....	16
3.6 Data Exploration and overview.....	16
3.6.1 Variable identification	16
3.6.2 Target variable.....	17
3.6.3 Predictor variables	17
3.6.4 Data Structure	18
3.7 Logistic regression analysis	18
3.8 Assumptions of the binary logistic regression	20
3.9 Survival Analysis	20
3.10 Hosmer-Lemeshow goodness of fit test for logistic regression	21
3.11 Hosmer-Lemeshow goodness of fit test for survival analysis.....	22
3.12 Multi-collinearity	22
3.13 Data Preparation, Presentation and Analysis procedures.....	22
3.14 Ethical Considerations.....	22
3.15 Chapter Conclusion.....	23
CHAPTER FOUR: DATA PRESENTATION, ANALYSIS AND DISCUSSION.....	24
4.0 Introduction.....	24
4.2 Descriptive statistics.....	24
4.3 Age distribution of the borrowers	26
4.3 Correlation tests.....	27
4.4 Test for multi-collinearity	28
4.5 Economic Indicators.....	29
4.6 Logistic Regression Analysis	29
4.7: Probability of default prediction formulae.....	30

4.8: Model efficiency test.....	30
4.9 Survival Analysis	31
4.9.1 Hazard curve.....	32
4.9.2 Model fitting.....	33
4.9.3 Model efficiency test	35
4.10 Comparison of the two models.....	36
4.11 Discussion of findings.....	38
4.12 Chapter Conclusion.....	39
CHAPTER 5: SUMMARY, RECOMMENDATIONS AND CONCLUSIONS	40
5.0 Introduction	40
5.1 Summary of findings.....	40
5.2 Contribution to the study.....	42
5.3 Recommendations	42
5.4 Credit Appraisal model	42
5.5 Consistent monitoring	43
5.6 Conclusions	43
5.7 Areas for further research.....	43
5.8 Chapter Conclusion	44
REFERENCES	45
APPENDICES	48

LIST OF FIGURES

Figure 4.1 Histogram for Age distribution	26
Figure 4.2 Correlation test	27
Figure 4.3 Curve for survival function	32
Figure 4.4 Hazard function curve	33
Figure 4.5 Logistic regression at current case	36
Figure 4.6 Survival analysis evaluation at current case	37

LIST OF TABLES

Table 3.1 Explanatory Variables.....	19
Table 4.2 Descriptive Statistics.....	24
Table 4.3 Multi-collinearity test	28
Table 4.4 Hosmer_Lemshow goodness of fit test.....	29
Table 4.5 Hosmer_Lemshow test	29
Table 4.6 Variables in the equation and their significance	30
Table 4.7 Efficiency classification of the model	31
Table 4.8 Values in the equation.....	34
Table 4.9 Classification table.....	35
Table 4.10 AUC Logistic regression	36
Table 4.11 AUC Survival Analysis	38

LIST OF ACRONYMS

RBZ	Reserve Bank of Zimbabwe
MFI	Micro-finance Institutions
GKCI	Gweru KCI
IMF	International Monetary Fund
FC	Financial Crisis
PD	Probability of Default
ALLL	Reserve for the loan losses
SPSS	Statistical Package for Social Sciences
ROC	Receiver Operating Characteristic

CHAPTER ONE: INTRODUCTION

1.0 Introduction

The first chapter of this study offers an overview of the problem being addressed and its roots, while also outlining the specific focuses of the research. This is done through sub-sections such as the problem statement, research objectives and questions or hypothesis, scope, significance, assumptions and limitations. This chapter is designed to provide readers with comprehensive information on the study. The section closes with a summary that condenses the information into a concise account. The research endeavors to address the critical issue of loan default prediction, a crucial aspect of credit risk assessment in the banking industry, by exploring the application of logistic regression and survival analysis techniques to contribute to the development of more accurate and effective predictive models, ultimately informing better decision-making processes in financial institutions.

1.1 Background to the Study

Less than 20% of the population in developing countries is served by the formal sector (Berenbach and Churchill, 1997; Robinson, 2001). Those who fall under the low-income earners and informal traders category are not able to access formal financial services. To address this, Microfinance Institutions (MFIs), commonly known as Innovative Financial Institutions, emerged to serve as a financing source for individuals and businesses in need of capital. The World Bank (WB) has estimated that about a billion households live on or below \$1.90 a day (WB, 2001). MFIs play a crucial role by providing small packages that are suitable for economically active persons (Zainal et al., 2019), making them vital to the economy.

The microfinance sector in Zimbabwe has recently faced significant challenges, particularly in the wake of the 2015 Mid-Term Monetary Policy Statement. At that time, at least 30 microfinance institutions closed due to capitalization struggles, decreasing the total number from

147 to 116 by the end of September 2015. This trend has continued, with increasing non-performing loans and a decline in investor confidence, ultimately leading to the current default. The need for effective risk management strategies and capitalization has become even more pressing, necessitating a fresh approach to ensure the survival and sustainability of microfinance institutions in the Zimbabwean economy.

The risk of loan default poses a considerable challenge to these institutions, threatening their financial stability and mission of poverty alleviation. As a result of, statistics shows that they have been a rapid increase in interest of researchers in developing accurate and robust predictive models to assess the credit risk of borrowers.

Microfinance institutions (MFIs) have emerged as a crucial player in the financial inclusion landscape, particularly in developing economies. These institutions provide a range of financial services, such as small loans, savings, and insurance, to individuals and small businesses who often lack access to traditional banking services. The primary objective of MFIs is to empower marginalized communities and foster economic development by extending credit to those who are typically considered high-risk borrowers.

However, the inherent nature of microfinance operations, characterized by small loan sizes, limited collateral, and high-risk clientele, poses significant challenges in assessing and managing credit risk. Accurate prediction of loan defaults is crucial for MFIs to maintain financial sustainability, maximize their outreach, and ensure the stability of the microfinance sector as a whole.

Conventional credit risk assessment methods, such as logistic regression, have been widely adopted by MFIs to predict the probability of loan defaults. Logistic regression models provide a robust framework for analyzing the relationship between borrower characteristics and the likelihood of default. These models have been extensively utilized in the microfinance context to identify the key determinants of loan repayment behavior and to develop predictive tools for credit decision-making (Masood & Srivastava, 2021; Xu & Liang, 2022).

While logistic regression has proven effective in many situations, it has certain limitations in the microfinance context. Logistic regression models assume that the risk of default is constant over the loan's lifetime, which may not always be the case. In reality, the risk of default can vary over

time due to changes in borrower characteristics, economic conditions, or other time-dependent factors.

Survival analysis, a statistical technique used to model time-to-event data, has emerged as a complementary approach to credit risk assessment in the microfinance industry. Survival analysis models can capture the dynamic nature of credit risk by incorporating time-varying covariates and providing insights into the timing and determinants of loan defaults (Hernandez-Mireles & Pedrosa, 2020). By considering the time dimension, survival analysis can enhance the predictive accuracy of credit risk models and provide MFIs with a more comprehensive understanding of their loan portfolio's performance.

The integration of logistic regression and survival analysis has been an area of growing interest in recent years, as researchers and practitioners seek to leverage the strengths of both approaches to improve credit risk management in the microfinance sector. This combined approach can offer MFIs a more nuanced and robust framework for assessing credit risk, enabling them to make more informed decisions, optimize their lending strategies, and ultimately enhance their overall financial sustainability and social impact.

1.2 Statement of the Problem

The accurate modeling of loan default probability is crucial for effective risk management and decision-making by microfinance institutions (MFIs). While logistic regression is commonly used, it fails to capture the dynamic nature of credit risk. Survival analysis can complement logistic regression to provide a more comprehensive understanding of loan default behavior. However, very few studies in the Zimbabwean microfinance sector have integrated these two approaches to assess default probabilities. This gap hinders MFIs' ability to develop robust credit risk assessment models and make informed decisions to enhance their financial sustainability and social impact. This study aims to address this gap by creating a comprehensive model that integrates logistic regression and survival analysis to predict loan default probability, thereby improving credit risk evaluation and decision-making in the microfinance industry in Zimbabwe.

1.3: Research objectives

This study seeks to:

1. To identify and examine the key drivers of loan default at KCI.
2. To create a predictive model that accurately forecasts loan default probability based on borrower and loan attributes.
3. To evaluate the performance of the developed statistical models in predicting loan default probability.

1.4 Research Questions

1. What are the key factors that influence the probability of loan default?
2. Which statistical models between logistic regression and survival analysis perform best in predicting loan default probability?
3. What is the predictive performance of the developed statistical models in terms of sensitivity, specificity, and area under the ROC curve (AUC-ROC)?

1.5 Scope of the Study

This study focuses on analyzing loan default probability within the microfinance sector in Zimbabwe. It aims to develop a comprehensive credit risk assessment model by integrating logistic regression and survival analysis techniques. The study utilized historical loan data from GKCI database and the Reserve Bank of Zimbabwe, covering the period from January 2020 to December 2023. The findings of this research provided valuable insights to microfinance institutions in Zimbabwe, enabling them to enhance their credit risk management practices and make more informed lending decisions. However, the generalizability of the results to other geographical contexts or microfinance markets may be limited and require further investigation.

1.6 Significance of the study

The significance of this study lies in its potential to enhance the decision-making capabilities of microfinance institutions (MFIs) in Zimbabwe. By integrating logistic regression and survival analysis techniques, the developed credit risk assessment model enabled MFIs to more accurately prediction of loan default probabilities. This, in turn, allows them to make more informed lending decisions, improving their financial sustainability, and safeguard their operations during periods of economic volatility. The findings of this research benefits MFI managers, credit officers, and risk management professionals by providing them with a robust framework in evaluating and managing credit risk. Additionally, the insights generated informs policy discussions and regulatory frameworks to strengthen the overall resilience of the microfinance sector in Zimbabwe

1.7 Assumptions of the study

In conducting this study on loan default probabilities, it is essential to establish a foundation of assumptions that guide the analysis and interpretation of results. These assumptions ensure the validity and reliability of the findings, allowing for meaningful conclusions to be drawn. The following assumptions underpin this research:

1. The probability of default for one loan applicant is not influenced by the default of another.
2. The relationships between the independent variables and loan default probabilities are linear and stable over time.
3. The data used for analysis is representative of the population of loan applicants and is free from significant biases.

1.8 Limitations of the Study

While this study provided valuable insights into loan default probabilities, it is essential to acknowledge the limitations that impacted the generalizability and applicability of the findings. The following limitations were encountered in this study:

1. The exclusion of certain variables in the models suggests that it will closely approximate the real phenomenon but not achieve an exact replication.
2. The study's reliance on historical data may not account for emerging trends or changes in loan applicant behavior, which could impact the model's predictive accuracy and generalizability to future scenarios.

1.9 Definition of terms

1. Loan Default: Loan default refers to the borrower's inability to repay the loan as per the agreed terms (Gup, 2011).
2. Logistic Regression: Logistic regression is a statistical method used to model the relationship between binary dependent variables, such as loan default, and a set of independent variables (Agresti, 2015).
3. Survival Analysis: Survival analysis is a statistical technique used to analyze the time it takes for an event, such as loan default, to occur (Klein & Moeschberger, 2003).
4. Credit Risk: Credit risk refers to the potential risk of financial loss due to the borrower's incapacity to repay a debt obligation (Basel Committee on Banking Supervision, 2000).
5. Microfinance Institutions (MFIs): Microfinance institutions are financial institutions that offer loans and other financial services to small businesses and individuals in the informal sector (Nguyen & Canh, 2020).
6. Loan Default- refers to the borrower's inability to repay the loan (Benton E Gup, 2012).
7. Logistic Regression- is a statistical method used to model the connection between binary dependent variables such as loan default (Agresti, A. 2015).
8. Survival Analysis- is a statistical technique used to analyze the time it takes for an event to happen, such as loan default (John P. Klein and Melvin L. Moeschberger, 2014).
9. Credit Risk- refers to the potential risk of financial loss because of the borrower's incapacity to repay a debt obligation (Principles for the Management of Credit Risk, October 2000).
10. Microfinance Institutions (MFIs) are financial institutions that offer loans to small businesses and the non-formal sector (Nguyen, B., Canh, N.P. 2008).

1.10 Conclusion

In conclusion, this chapter has laid the foundation for the study by providing a comprehensive overview of the research topic, objectives, and key definitions. With this solid groundwork in place, the next chapter will delve into the literature review, exploring existing research on loan default, logistic regression, survival analysis, credit risk, and microfinance institutions, setting the stage for the analysis and findings to come.

CHAPTER TWO: LITERATURE REVIEW

2.0 Introduction

The aim of this section is to examine the theoretical and empirical studies that various authors have conducted in relation to the problem under investigation. The purpose of these reviews is to provide a critical analysis that aligns with the study's purpose and direction. This chapter provides an in-depth explanation of the problem and a thorough discussion of the research methods to be used.

2.1 Theoretical Literature

The proposed study is grounded in two primary theoretical frameworks: the credit risk management theory and the information asymmetry theory, both of which are widely recognized in the microfinance and banking literature.

2.2 Credit Risk Management Theory

The credit risk management theory emphasizes the importance of effectively identifying, measuring, and mitigating credit risk to ensure the financial sustainability of financial institutions (Basel Committee on Banking Supervision, 2011). In the context of microfinance, credit risk management is crucial as MFIs typically lend to underserved and higher-risk borrowers (Muriu, 2011). Robust credit risk assessment models that can accurately predict loan default probabilities are essential for MFIs to make informed lending decisions, price their products accordingly, and allocate capital efficiently (Dionne, 2013; Bouteille & Coogan-Pushner, 2014).

The credit risk management theory has been extensively applied in the microfinance sector, with studies highlighting the need for MFIs to develop comprehensive risk management frameworks to address the unique challenges they face. For example, Zamore et al. (2018) found that MFIs with more robust credit risk management practices were better able to withstand financial shocks and maintain their financial stability. Similarly, Marakkath and Attuel-Mendes (2015) emphasized the importance of integrating credit risk assessment with other risk management

strategies, such as portfolio diversification and risk-based pricing, to enhance the overall resilience of MFIs.

Credit risk management is a critical aspect of ensuring the sustainability and success of microfinance institutions (MFIs). The theoretical literature on credit risk management in microfinance draws heavily from the broader banking and finance literature, but also incorporates insights specific to the microfinance context.

At the core of credit risk management theory in microfinance is the need to effectively assess, monitor, and mitigate the risks associated with lending to underserved, often low-income borrowers. Seminal work by Stiglitz and Weiss (1981) highlighted how information asymmetries between lenders and borrowers can lead to adverse selection and moral hazard, ultimately undermining the ability of financial institutions to accurately price and manage credit risk.

In response to these challenges, microfinance theorists have proposed a range of credit risk management strategies tailored to the microfinance context. These includes:

1. Relationship lending and dynamic incentives: MFIs can build long-term relationships with borrowers, using mechanisms such as progressive lending (increasing loan sizes upon successful repayment) to incentivize repayment and gather valuable information about the borrower's creditworthiness over time (Armendáriz & Morduch, 2010).
2. Group-based lending and social collateral: By organizing borrowers into groups and making them jointly liable for each other's loans, MFIs can leverage social capital and peer monitoring to enhance credit risk management (Ghatak & Guinnane, 1999).
3. Comprehensive credit risk assessment: MFIs can employ rigorous screening and assessment procedures that consider not only the financial characteristics of the borrower, but also their broader socioeconomic context, livelihood strategies, and vulnerability factors (Ngo et al., 2021).
4. Diversification and portfolio management: MFIs can diversify their loan portfolios across different geographic regions, economic sectors, and borrower profiles to mitigate concentration risks and enhance the overall resilience of their credit operations (Bogan, 2012).

5. Innovative credit risk management tools: MFIs can explore the use of technology-enabled solutions, such as credit scoring models, digital credit assessments, and data analytics, to enhance the accuracy and efficiency of their credit risk management practices (Cull et al., 2018).

2.3 Information Asymmetry Theory

The information asymmetry theory posits that borrowers often have more information about their creditworthiness and repayment capacities than lenders, leading to adverse selection and moral hazard problems (Stiglitz & Weiss, 1981). In the microfinance sector, this information asymmetry is more pronounced due to the lack of formal credit histories and financial records for many borrowers (Hermes & Lensink, 2007). Integrating both logistic regression and survival analysis techniques can help MFIs overcome these information asymmetries and develop more accurate credit risk assessment models (Gašparienė & Kartašova, 2016; Calabrese & Osmetti, 2015).

The information asymmetry theory has been extensively applied in the microfinance literature, with researchers exploring various strategies to mitigate the effects of information asymmetry. For instance, Galema et al. (2018) found that the use of group lending and dynamic incentives, such as progressive lending, can help reduce information asymmetries and improve repayment rates. Similarly, Cull et al. (2014) highlighted the importance of credit information sharing and the development of credit bureaus in improving the ability of MFIs to assess the creditworthiness of borrowers.

By drawing on these two well-established theoretical frameworks, this study aims to create a comprehensive credit risk assessment model that can enhance the decision-making capabilities of MFIs in Zimbabwe, thereby improving their financial sustainability and social impact.

Information asymmetry is a central theoretical concept in the microfinance literature, as it helps to explain the unique challenges faced by MFIs in assessing and managing credit risk. The seminal work of Stiglitz and Weiss (1981) on information asymmetry in credit markets laid the foundation for understanding how these information gaps can lead to adverse selection and moral hazard, ultimately undermining the ability of lenders to accurately price and manage credit risk.

In the microfinance context, information asymmetry is particularly pronounced due to the lack of formal credit histories, collateral, and other traditional creditworthiness indicators among the

target population of low-income and often financially excluded borrowers (Armendáriz & Morduch, 2010). This information gap makes it difficult for MFIs to accurately assess the risk profile of potential borrowers, leading to higher transaction costs and greater uncertainty in the lending process.

To address the challenges posed by information asymmetry, microfinance theorists have proposed several strategies:

1. **Group-based lending and joint liability:** By organizing borrowers into groups and making them jointly liable for each other's loans, MFIs can leverage the power of social networks and peer monitoring to gather information about borrowers and incentivize repayment (Ghatak & Guinnane, 1999).
2. **Dynamic incentives and relationship lending:** MFIs can build long-term relationships with borrowers, using mechanisms such as progressive lending and the promise of future access to credit to gather information about borrowers' creditworthiness and encourage repayment (Armendáriz & Morduch, 2010).
3. **Comprehensive credit assessment:** MFIs can employ rigorous screening and assessment procedures that consider not only the financial characteristics of the borrower, but also their broader socioeconomic context, livelihood strategies, and vulnerability factors, to better understand their creditworthiness (Ngo et al., 2021).
4. **Innovative data and technology solutions:** MFIs can leverage digital platforms, alternative data sources, and advanced analytics to gather more granular information about borrowers and enhance their ability to assess and manage credit risk (Cull et al., 2018).

By drawing on these theoretical insights, MFIs can develop more effective strategies to bridge the information gaps that contribute to credit risk, ultimately enhancing their ability to provide responsible and sustainable financial services to underserved populations.

2.4 Empirical Literature

The empirical literature on credit risk assessment and management in the microfinance industry has grown considerably in recent years, with researchers exploring a variety of factors that

influence loan default and developing sophisticated credit risk models to enhance the decision-making capabilities of microfinance institutions (MFIs).

2.5 Borrower-Level Factors

Several studies have examined the impact of borrower-level characteristics on loan repayment performance. For instance, Roslan and Karim (2009) found that factors such as gender, marital status, and household size significantly affected loan repayment rates in Malaysia. Similarly, Nawai and Shariff (2013) identified education level, income, and loan size as key determinants of loan default in Malaysia. More recently, Bena and Jiang (2021) analyzed data from a leading Chinese MFI and concluded that borrower age, gender, and entrepreneurial experience were significant predictors of loan default.

Recent research by Ahuja and Karam (2021) explored the factors affecting loan defaults in India. The study found that economic conditions, borrower demographics, loan characteristics, and underwriting practices all played a role in loan defaults in the country. By understanding these factors, lenders can take steps to mitigate the risk of loan default and ensure that borrowers are able to repay their loans

2.6 Institutional Factors

Researchers have also investigated the role of institutional factors in credit risk management. Godquin (2004) examined the influence of group lending, dynamic incentives, and monitoring on loan repayment rates in Bangladesh, finding that these factors were positively associated with higher repayment performance. Likewise, Ahlin and Townsend (2007) analyzed data from Thailand and found that MFI characteristics, such as financing structure and operational efficiency, were important determinants of loan repayment rates.

2.7 Combining Techniques

To enhance the accuracy of credit risk assessment, researchers have explored the use of hybrid models that integrate multiple analytical techniques. For example, Gašparienė and Kartašova (2016) combined logistic regression and survival analysis to predict loan defaults in Lithuania, while Calabrese and Osmetti (2015) used the generalized extreme value regression model to assess credit risk for small and medium-sized enterprises in Italy. More recently, Delbanco et al.

(2021) integrated machine learning algorithms with traditional statistical methods to develop a robust credit risk assessment model for a leading MFI in India.

2.8 Contextual Factors

The importance of contextual factors, such as macroeconomic conditions and regulatory environment, has also been highlighted in the empirical literature. Mwangi and Sichei (2011) investigated the impact of macroeconomic variables on loan repayment rates in Kenya, finding that factors like inflation and interest rates significantly influenced loan default. Likewise, Ahlin et al. (2011) analyzed data from 73 countries and concluded that the legal and regulatory framework, as well as the level of competition in the microfinance sector, were key determinants of MFI performance.

By synthesizing the findings from this growing body of empirical research, the proposed study aims to develop a comprehensive credit risk assessment model that can effectively inform the decision-making processes of MFIs in Zimbabwe, thereby enhancing their financial sustainability and social impact.

2.9 Chapter Conclusion

The comprehensive literature review has provided a solid theoretical and empirical foundation for understanding the key determinants of credit risk in the microfinance industry. The analysis has revealed the multidimensional nature of loan repayment behavior, highlighting the importance of borrower characteristics, institutional policies, and contextual factors. Building on these insights, the next chapter will outline the methodology employed in this study to develop a robust credit risk assessment model that integrates these interrelated elements. This innovative framework aims to equip microfinance institutions in Zimbabwe with enhanced decision-making capabilities to improve their financial sustainability and social impact.

CHAPTER THREE: RESEARCH METHODOLOGY

3.0: Introduction

This chapter is dedicated to describing the data utilized in the study. It explains how the data was analyzed and manipulated to achieve optimal results that address the research questions. Additionally, the researcher provides a detailed explanation of the methods employed and justifies the use of specific methods where applicable.

3.1 Research Design

This study examined the factors influencing loan default and the timing of default events using a combination of logistic regression and survival analysis techniques. The research was conducted using data collected from the KCI database and the Reserve Bank of Zimbabwe, focusing on non-performing loans. The dataset covered a specific time period from January 2020 to December 2023, included 789 randomly selected loan accounts. The data encompassed relevant information on borrower characteristics, loan attributes, and macroeconomic indicators, as well as the occurrence and timing of loan defaults. The study considered various predictor variables, including borrower characteristics such as age, source of income, and credit status, as well as loan features like loan amount, loan grade, and repayment term.

The primary dependent variables in this study were the binary outcome of loan default (default or non-default) and the time until default, which was analyzed using logistic regression and survival analysis, respectively. The logistic regression model was used to predict the probability of loan default based on the selected predictors. Survival analysis, or time-to-event analysis, will then be employed to model the duration until loan default, taking into account factors like payment history. By combining the outcomes of the logistic regression and survival analysis, the study aims to provide a comprehensive framework for evaluating the probabilities and timing of loan defaults. This approach will contribute to a deeper understanding of the key determinants of loan default and the temporal dynamics involved. The data analysis involved thorough data preparation, including handling missing values, outliers, and the creation of any necessary derived variables. Diagnostics and model validation procedures were performed to ensure the

reliability and robustness of the statistical models. The expected outcomes of this research included insights into the primary drivers of loan default and the ability to predict the timing of default events. These findings have significant implications for financial institutions, informing their loan underwriting processes, risk management strategies, and portfolio optimization efforts. Additionally, the research contributed to the existing literature by advancing the understanding of loan default dynamics in the specific context of the KCI database and the Reserve Bank of Zimbabwe.

3.2 Data Sources

In this study, the researcher collected loan portfolio data from KCI Financial Institution. A sample of 789 distinct loan accounts was obtained for the analysis. These observations were extracted from the KCI ANALYSIS BOOK and imported it into an excel spreadsheet for further analysis, starting from January 2020 to December 2023. The Analysis Book is an all-encompassing compilation of loan data. The researcher took measures to gather all necessary information, including reaching out to previous clients for any missing data, while adhering to strict confidentiality and privacy regulations throughout the entire data collection process.

3.3 Target population and sampling methods

The researcher conducted a comprehensive study on micro-financial institutions, with a specific emphasis on KCI Management Consultants as it is one of the prominent microfinance institutions in Zimbabwe, operating in a provincial town context where access to formal financial services is limited for many individuals and small businesses. KCI is a privately owned microfinance institution situated in Gweru, Midlands's province. The province covers an area of 49 166 square kilometers with a population of 1621 656 people (Zhujiworld.com, July 2022). Gweru has a population of 146 073 people and is the capital of the province (World Population Review). While the primary focus is on the KCI Gweru Branch, there was a portion of the KCI Gweru branch population comprises clients residing outside Gweru for example Shurugwi.

The study endeavored to engage the population of individuals in the Midlands province utilizing KCI's services, including both men and women who are eligible to apply for a loan. To ensure inclusiveness, on creating the sample dataset 789 loan accounts were selected for the period of

January 2020 to December 2023. The goal was to ensure that the findings represent the views and experiences of the diverse population that KCI serves.

3.4 Research Instruments

The research instrument used in this study was the Analysis Book of loan borrowers, which provided a comprehensive dataset of loan characteristics from January 2020 to December 2023. This archival data source offered a wealth of information, including loan amount, debt-to-income ratio, repayment terms, and purpose of the loan. By leveraging this existing data, the study was able to model loan default probabilities and analyze various loan characteristics. The Analysis Book data served as the sole data source for this research, providing a robust and reliable foundation for the study's findings.

3.5 Methods for data collection

This study relied solely on secondary data sources to fulfill its research objectives. A vast array of secondary data sources were utilized, including but not limited to statistical reports on non-performing loans published by the Reserve Bank of Zimbabwe. These reports provided valuable insights into the trends and patterns of non-performing loans in the banking sector. Additionally, the researcher also drew upon the extensive database of KCI Management Consultants, which offered a wealth of information on the subject matter. By leveraging these secondary data sources, the researcher was able to gather a comprehensive understanding of the research topic, sans the need for primary data collection. This approach proved particularly useful in expediting the research process, as it eliminated the time-consuming and resource-intensive requirements associated with primary data collection. Furthermore, secondary data sources provided a robust foundation for analysis, enabling the researcher to draw meaningful conclusions and make informed recommendations.

3.6 Data Exploration and overview

Before proceeding with the statistical analysis, it is important to explore and understand the structure of the data. This section examined the variables, their data types, and the overall data structure.

3.6.1 Variable identification

The key variables included:

1. Borrower age: the age of the borrower at the time of loan default (discrete, ranges from 20 to 56; 20 being the youngest and 56 being the eldest),
2. Income group: the annual income of the borrower {categorical, ranges from 1 to 3; identified as low, medium and high respectively},
3. Employment length: employment length in years {categorical, ranges from 0 to 10; 0 being employment less than 1 year and 10 represents employment more than 10 years respectively},
4. Loan amount: the amount of the loan requested or granted to the borrower { discrete, ranges from 1000 to 35000},
5. Loan grade: Lc assigned grade (categorical; ranges from 1 to 7; identified as grade A to G respectively),
6. Household ownership status: home ownership status of the borrower {categorical: ranges from 1 to 3; identified as own, rent and mortgage}
7. Loan purpose: purpose of the loan applied for by client {categorical: ranges from 1 to 13 identified as credit card, car, small business, other, wedding, debt consolidation, home improvement, major purpose, medical , moving, vacation, house, renewable energy respectively}.
8. Loan id: unique identifier for each loan(discrete)
9. Debt to income ratio: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations. (discrete, ranges from 0.32 to 29.83)

3.6.2 Target variable

The target variable of interest was the binary loan default indicator (0= no default, 1=default).

This would be the key outcome that the analysis aimed to model and predict.

3.6.3 Predictor variables

The predictor variables in this study were:

1. Borrower Age
2. Loan Amount
3. Loan Purpose
4. Income

5. Debt to income ratio
6. Loan assigned grade
7. Home ownership status
8. Employment length

These variables had an influence on the probability of loan default and were used in the statistical modeling.

3.6.4 Data Structure

The dataset consists of 789 loan records with the variables mentioned above. The data types of the variables were a mix of categorical and discrete, which required appropriate handling during the analysis.

Exploring the data in this manner gave us a better understanding on the variables and their relationships, which informed subsequent statistical modeling and analysis.

3.7 Logistic regression analysis

To conduct the logistic regression analysis, the researchers utilized SPSS version 16 software and employed the forward selection method to identify the most significant variables. This type of analysis is best suited for binary outcomes, where the dependent variable can take on one of two possible values, as stated by Wuensch (2014). The objective of the study was to predict the likelihood of loan default. The explanatory variables considered in the analysis could be nominal, ordinal, or interval in nature. It is important to note that regression analysis does not require predictor variable distributions to be specified, according to Burns and Burns (2018).

$$Y = \beta_0 + \sum_{n=1}^8 \beta_n X_n$$

Where the dependent variable $Y =$ either 0 where there is no default or 1 when the borrower defaulted with probabilities $(1-p)$ or p respectively and in the case of this thesis p is the probability of a client defaulting.

β_0 is the y intercept,

β_1 is the Beta coefficients of the respective variable.

The Explanatory Variables are as shown on Table 3.1 below,

Table 3.1 Explanatory Variables

	Dependent Variable
X ₁	Age
X ₂	Employment length
X ₃	Home ownership status
X ₄	Purpose of loan
X ₅	Loan grade
X ₆	Debt to income ratio
X ₇	Source of income
X ₈	Loan amount

Link function: logit

The logistic function is a popular, effective, and clear method for categorizing things as good or negative. This is a function that takes in client information and returns the probability of default.

$$\text{Log (odds)} = \log\left(\frac{p}{1-p}\right) = \beta^0 + \sum_{i=1}^8 \beta_i X_i$$

Or more equivalent to,

$$p = \frac{\text{odds}}{1 + \text{odds}}$$

$$p = \frac{\exp(\beta_0 + \beta_1 \cdot x_1 + \dots + \beta_8 x_8)}{1 + \exp(\beta_0 + \beta_1 \cdot x_1 + \dots + \beta_8 x_8)}$$

In the preceding equation,

1. p represents the probability of default,
2. X_i is the explanatory variable, and
3. β_i is the regression coefficient for the explanatory variable i .

The loan status of each existing data point specifies whether the client defaulted ($p=1$ or $p=0$). The goal is to determine the coefficients $\beta_0 \dots \dots \beta_8$ so that the model's probability of default matches the observed probability of default. Maximum likelihood estimation is used to determine the values of β_i

3.8 Assumptions of the binary logistic regression

1. The binary logistic regression model makes the assumptions listed below,
2. The dependent variable needs to be binary.
3. It assumes that the predictor variables are independent.
4. It presupposes that there is no multi-collinearity among the independent variables
5. It presupposes linearity of independent variables and log odds.

3.9 Survival Analysis

Survival Analysis, a statistical technique utilized in this thesis, examines data by measuring the amount of time needed for a specific event of interest to occur. The primary focus of this study is to trace the subjects for a set period and observe events that relate to the research. This thesis centers on examining the occurrence of loan defaults, and the duration between occurrences can be measured in weeks, months, or years. However, the time intervals in this research are measured in days. To model survival analysis, T represents the duration from the loan issuance to the moment of default, and $F(t)$ represents the cumulative density function. Or, in other words, it is the likelihood that the loan will default before or at time t . In mathematical terms, $F(t) = P(T \leq t)$. Thus, survival function is interpreted as the probability of observing the event after time, $T=t$; $S(t) = P(T > t) = 1 - F(t)$.

The hazard function is ideal for continuous survival data. It is defined as,

$$h(t) = \lim_{\Delta t \rightarrow 0} \left(\frac{P(t \leq T < T + \Delta t | T \geq t)}{\Delta t} \right)$$

The hazard function assesses the probability of an event occurring at a particular time, while the survival probability calculates the likelihood of a loan remaining in good standing beyond a specific time frame. The survival probability is represented by $S(t|\tau) = p(T_i > \tau)$, where $i = 1, 2, \dots$ and $\tau = 1, 2, \dots, M$ indicating the number of days. The hazard function at a given time, t , is the probability of default over τ ($T_i = \tau$), given that there has been no default before that time ($T_i \geq \tau$). Here, the probability of default is expressed as a conditional probability. The hazard function, $h(t|\tau) = p(T_i = \tau | T_i \geq \tau)$ is dependent on the absence of default up to time τ and is unique to each loan based on its specific characteristics. Survival Analysis models the distribution of default times, enabling one to forecast the probability of default within a given time frame. The probability of default can be presented using the hazard function and a logit link function, similar to a logistic regression model. The model is represented as follows.

Logit $h(t) = \log\left(\frac{h(t)}{1-h(t)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_8 x_8$, which can be simplified to,

$$h(t) = \log \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_8 x_8)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_8 x_8)}, \text{ where } x_1, x_2, \dots, x_8 \text{ are the determinants of loan default.}$$

3.10 Hosmer-Lemeshow goodness of fit test for logistic regression

Before using an assumed model for prediction or drawing conclusions, it should be validated to ensure it is appropriately defined. Data should align with the model's assumptions. Allison (2014) sought clarification on how one may see or know whether the model fits the data. The researcher used the Hosmer-Lemeshow goodness of fit test to ensure that the logistic regression fits the data.

H_0 : the current model fits well

H_1 : the current model does not fit well

The test was conducted at 5% level of significance.

3.11 Hosmer-Lemeshow goodness of fit test for survival analysis

Data should align with the model's assumptions. Allison (2014) sought clarification on how one may see or know whether the model fits the data. The researcher used the Hosmer-Lemeshow goodness of fit test to ensure that the survival analysis fits the data.

H_0 : the current model fits well

H_1 : the current model does not fit well

The test was conducted at 5% level of significance.

3.12 Multi-collinearity

Multi-collinearity is a phenomenon that occurs when two or more predictors in a regression model are interrelated. These predictors can be highly or moderately associated with each other. In such a scenario, multi-collinearity can reduce the relevance of an independent variable, leading to misleading results in the analysis. However, during our evaluation of the association between loan default determinants and default status, we found no evidence of multi-collinearity. If we had identified this issue, we could have taken corrective measures such as removing one of the strongly associated variables, to ensure the accuracy of the analysis.

3.13 Data Preparation, Presentation and Analysis procedures

After collecting data from GKCI, data management techniques were implemented to minimize the likelihood of prospective bias. The data was cleansed and altered where appropriate, and visualization methods were employed where needed to comprehend the trends and critical characteristics of the data. These data management techniques are vital for obtaining accurate, reliable, and impartial results during the data analysis stage of the research. The researcher performed a model analysis on both logistic regression and survival analysis to account for the requirement of comprehensive explanations for loan default dynamics in this study.

3.14 Ethical Considerations

To prioritize the anonymity of clients, unique identification numbers were assigned to their data in this study. The data was collected voluntarily from GKCI, and approval was obtained from the KCI regional manager for academic purposes. However, the data was not altered in any way to obtain dependable results while maintaining GKCI's reputation in the process. The majority of the data utilized in the study was quantitative, which helps to safeguard the participants' information.

3.15 Chapter Conclusion

This section of the study described how the data was utilized and managed. It also presented the formulation of a model, along with an explanation of the model analysis techniques and methods that was employed in the study. Generally, this chapter explained all the methods that were implemented in the subsequent chapter

CHAPTER FOUR: DATA PRESENTATION, ANALYSIS AND DISCUSSION

4.0 Introduction

In this chapter, the focus is to discuss and analyze the findings of the study. The results are presented in tables, cross tabulations, and graphs, which are thoroughly explained and discussed. There was a compilation of the results from the SPSS and analyzed a Logistic Regression Model and Survival analysis to understand the predictive power of each model. There was also a presentation and discussion of several parameters, and a comprehensive analysis of various strategies to mitigate loan default.

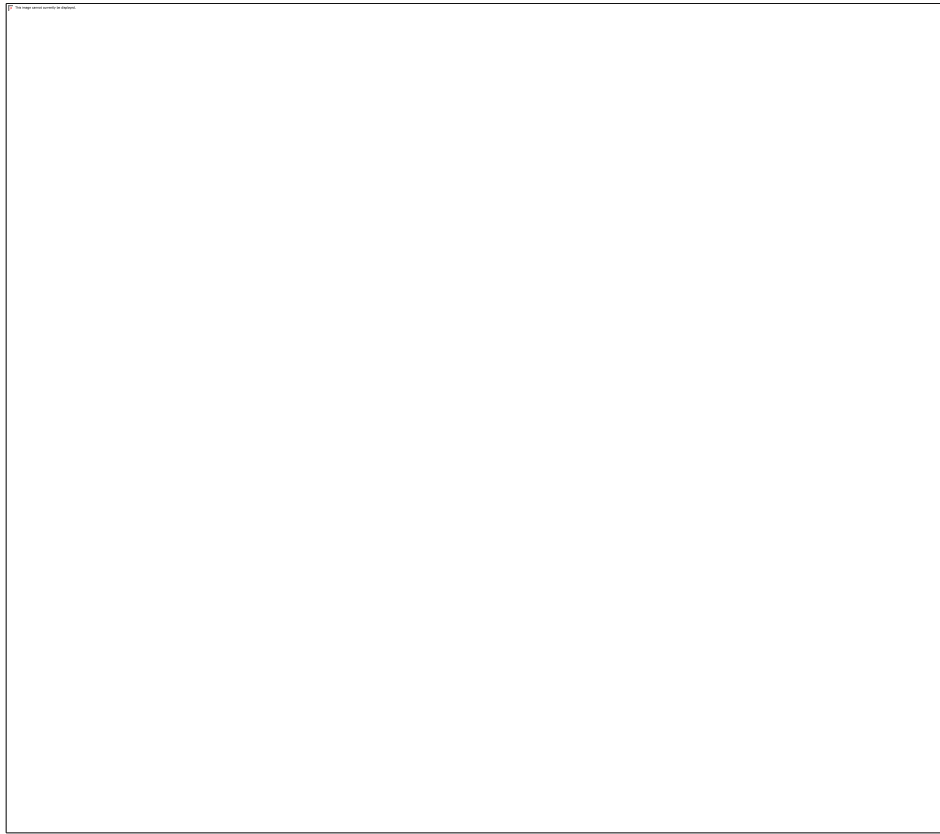
4.2 Descriptive statistics

Table 4.2 Descriptive Statistics

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Age	789	20	56	35.00	8.040
emp_length_int	789	.00	10.00	5.6755	3.33503
home_ownership_cat	789	1	3	1.87	.956
income_cat	789	1	3	1.13	.369
loan_amount	789	1000	35000	1.34E4	8039.654
purpose_cat	789	1	13	4.98	2.427
Lc	789	1	7	2.75	1.420
Dti	789	.3200	29.8300	1.442162E 1	6.2756773
loan_default_status	789	0	1	.16	.370
Valid N (listwise)	789				

Table 4.2 above summarizes eight variables from 789 borrowers' loan accounts, with various characteristics related to loan applications or credit profiles. The age range of 20 to 56, with an average age of 35, suggests that most individuals are likely established in their careers and have some level of financial stability. Employment length ranges from 0 to 10 years, with an average of 5.68 years, indicating that many individuals have some level of work experience. Home ownership categories and income categories are also included, with average values of 1.87 and 1.13, respectively, which may suggest that most individuals are renters or have lower to moderate incomes. Loan amounts range from \$1,000 to \$35,000, with an average of \$13,400, indicating that many individuals are seeking personal loans or smaller lines of credit. The purpose of the loans varies, with an average category value of 4.98, which may suggest that most individuals are applying for loans for reasons like credit card debt or personal expenses. Additionally, the loan-to-value ratio (Lc) ranges from 1 to 7, with an average of 2.75, indicating that individuals are borrowing a moderate amount compared to the value of their collateral. Finally, the debt-to-income ratio (Dti) ranges from 0.32 to 29.83, with an average of 14.42, which may indicate that many individuals have a relatively high debt burden compared to their income. Overall, this dataset provides a snapshot of the characteristics of these individuals, which can be useful for understanding credit profiles and loan applications.

Figure 4.1 Histogram for Age distribution



4.3 Age distribution of the borrowers

The age data follows a normal distribution with an average age of 35.00 years .The youngest participant being 20 and oldest participant being 56 years. As depicted in Figure 4.1, a significant proportion of respondents, accounting for 42.20% belonged to the 27-41 years age group. This distribution shares similarities with that of Dzingai (2014), who studied agro-based credit schemes in the tobacco industry. In that study, the average age of a farmer was 41.45 years, with the youngest participant being 21 years and the oldest being 65years old. The majority of respondents fell within the 31-40 years age category.

4.3 Correlation tests

Figure 4.2 Correlation test

		age	emp_leng th_int	home_ow nership_c at	income_c at	loan_amo unt	purpose_ cat	Lc
age	Pearson C	1						
emp_lengt	Pearson C	0.046788	1					
home_own	Pearson C	-0.07479	0.177345	1				
income_ca	Pearson C	0.858274	0.050147	0.184277	1			
loan_amou	Pearson C	0.776353	0.139503	0.170439	0.284609	1		
purpose_c	Pearson C	0.245002	-0.04144	0.055632	0.057313	0.005153	1	
Lc	Pearson C	-0.06137	0.037568	-0.0301	0.079393	0.38796	-0.02782	1
dti	Pearson C	0.705469	0.084855	-0.09365	-0.20168	0.050721	-0.07371	0.122656

The correlation matrix depicted in Figure 4.2 above shows the correlation between independent and predictor variables. It is clear that some of the independent variables are at most moderately marginally correlated. For example Age have a moderately positive correlation with employment length (0.046788) and purpose of loan (0.245002). Also the negative shows the inverse proportionality, for example Age and house ownership status are weakly negatively related (-0.07479) indicating younger borrowers have a better house ownership status than the older ones but the impact is small to be considered significant. Age and debt to income ratio are fairly strongly correlated (0.705469) this shows a direct relation and the impact is fairly strong showing that there is a better correlation.

4.4 Test for multi-collinearity

For data to fit for logistic regression model, there is need for multi-collinearity and be dealt with when found. Multi-collinearity can be detected using Variance Inflation Factor (VIF). If the VIF value is less than 10 or equal to 10 it means that no severe multi-collinearity exists in the model. As shown in Table 4.2 below we can see the last VIF column, as per the bench mark all the values are less than 10. Hence there is no multi-collinearity in the model.

Table 4.3 Multi-collinearity test

Coefficients ^a									
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1 (Constant)	.080	.089		.900	.368	-.094	.254		
Age	.000	.002	-.013	-3.87	.699	-.004	.002	.985	1.015
emp_length_int	.000	.004	.002	.062	.951	-.007	.008	.942	1.062
home_ownership_cat	-.004	.014	-.010	-.289	.773	-.031	.023	.903	1.108
income_cat	-.102	.037	-.101	2.739	.006	-.174	-.029	.857	1.166
loan_amount	-1.011E-6	.000	-.022	-.558	.577	.000	.000	.754	1.326
purpose_cat	.005	.005	.035	1.017	.309	-.005	.016	.988	1.012
Lc	.076	.010	.290	7.726	.000	.056	.095	.827	1.210
Dti	.000	.002	.006	.182	.856	-.004	.005	.921	1.086

a. Dependent Variable: loan default

4.5 Economic Indicators

According to the Zimbabwe Labor Market Diagnostic Analysis-ILO, the unemployment rate in Zimbabwe during the data collection period (January 2020 – December 2023) averaged around 4.5% officially, and its estimated GDP was \$56 billion USD.

4.6 Logistic Regression Analysis

Table 4.4 Hosmer_Lemshow goodness of fit test

Contingency Table for Hosmer and Lemeshow Test

		loan_condition_cat = 0		loan_condition_cat = 1		Total
		Observed	Expected	Observed	Expected	
Step 1	1	76	75.846	3	3.154	79
	2	75	73.900	4	5.100	79
	3	70	72.734	9	6.266	79
	4	69	71.098	10	7.902	79
	5	72	69.999	7	9.001	79
	6	68	68.020	11	10.980	79
	7	70	65.188	9	13.812	79
	8	61	62.292	18	16.708	79
	9	56	56.515	23	22.485	79
	10	43	44.406	35	33.594	78

Table 4.5 Hosmer_Lemshow test

Step	Chi-square	Df	Sig.
1	4.956	8	.762

Table 4.5 above shows p-value for the model of 0.762, which is higher than the significance level of 0.05. Hence, we fail to reject the null hypothesis, and the conclusion is that the model fits well with the data and is significant statistically. This reveals that the model is a dependable and suitable fit for the data.

4.7: Probability of default prediction formulae

Table 4.6 Variables in the equation and their significance

		Variables in the Equation					95.0% C.I.for EXP(B)		
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1 ^a	Age	-.005	.013	.132	1	.716	.995	.971	1.021
	emp_length_int	.000	.032	.000	1	.984	.999	.940	1.063
	home_ownership_cat	-.051	.112	.202	1	.653	.951	.763	1.185
	income_cat	-1.077	.403	7.130	1	.008	.341	.154	.751
	loan_amount	.000	.000	.547	1	.460	1.000	1.000	1.000
	purpose_cat	.042	.044	.921	1	.337	1.043	.957	1.136
	Lc	.551	.078	50.025	1	.000	1.735	1.489	2.021
	Dti	.007	.017	.150	1	.699	1.007	.973	1.041
	Constant	-2.050	.778	6.944	1	.008	.129		

a. The emp length and the loan amount will be left out of the equation since it is relatively too small to be included thus: x_1 represents age, x_2 represents home_ownership_cat, x_3 represents income_cat, x_4 represents, purpose_cat, x_5 represents Lc, and x_6 represents dti.

$$= \frac{\exp(-2.050 - 0.005x_1 - 0.051x_2 - 1.077x_3 + 0.042x_4 + 0.551x_5 + 0.007x_6)}{1 + \exp(-2.050 - 0.005x_1 - 0.051x_2 - 1.077x_3 + 0.042x_4 + 0.551x_5 + 0.007x_6)}$$

This probability of default prediction formula can be used to predict the default probability of the future client or applicant using their respective variables

4.8: Model efficiency test

Table 4.7 Efficiency classification of the model

Classification Table^a

Observed			Predicted		Percentage Correct
			loan_condition_cat		
			0	1	
Step 1	loan_condition_cat	0	652	8	98.8
		1	121	8	6.2
	Overall Percentage				83.7

Table 4.7 above serves a specific purpose of evaluating the classification effectiveness of the model. It also serves as a valuable method for assessing how well the model matches the data. According to the table, 652 did not default and were correctly classified showing 98.8% and it shows the probability of error of error of 1.8% (1 – 98.8%) which is an acceptable margin of error with an overall percentage correct of 83.7% and thus the model presents the data well.

4.9 Survival Analysis

By transforming the cumulative distribution function, we obtain the survival function $S(t)=1-F(t)$. This function represents the likelihood of surviving beyond a particular time point t , which is equivalent to the probability of $\Pr(T>t)$ of default not happening. The analysis showed that most defaults happened within 36 months, with many occurring in the 24 to 36-month range. This finding may possibly be attributed to adjustable interest rates resulting in lower monthly payments.

Figure 4.3 Curve for survival function



4.9.1 Hazard curve

Survival Analysis primarily revolves around modeling the hazard rate, denoted as $h(t)$. The hazard rate specifically represents the instantaneous failure rate, such as the rate of defaults, at any given time point t . It disregards the cumulative hazard up to that point. In the observed data, the hazard rate exhibits a consistent upward trend until the 24-month mark, at which point it experiences a sudden sharp increase. Following this abrupt rise, the hazard rate gradually continues to increase until another sudden rise occurs.

Figure 4.4 Hazard function curve



4.9.2 Model fitting

The model that is applied is as follows,

Logit $h(t) = \log \frac{h(t)}{1-h(t)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_8 x_8$, which can be simplified to,

$h(t) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_8 x_8)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_8 x_8)}$, where x_1, x_2, \dots, x_8 are variables described in chapter three of this project.

Table 4.8 Values in the equation

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	.000	.011	.007	1	.932	.999	.977	1.021
emp_length_int	-.009	.028	.095	1	.758	.992	.939	1.047
home_ownership_cat	-.073	.099	.550	1	.458	.929	.765	1.128
income_cat	-.793	.366	4.697	1	.030	.452	.221	.927
loan_amount	.000	.000	.476	1	.490	1.000	1.000	1.000
purpose_cat	.035	.038	.849	1	.357	1.035	.962	1.114
Lc	.406	.064	40.622	1	.000	1.501	1.325	1.700
Dti	.003	.015	.047	1	.828	1.003	.974	1.033

$$h(t) = \frac{\exp(-0.009x_1 + 0.073x_2 + \dots - 0.003x_8)}{1 + \exp(-0.009x_1 + 0.073x_2 + \dots - 0.003x_8)}$$

4.9.3 Model efficiency test.

Table 4.9 Classification table

Classification Table^a

Observed			Predicted		Percentage Correct
			loan_condition_cat		
			0	1	
Step 1	loan_condition_cat	0	652	8	99.0
		1	121	8	6.2
	Overall Percentage				83.8

Table 4.9 has a specific purpose of evaluating the classification performance of the model. It also serves as a valuable tool for assessing how well the model aligns with the data. According to the table, out of the total cases, 411 were actually classified as default, resulting in a success rate of 91.9%. Consequently, the model effectively represents the data.

4.10 Comparison of the two models.

Figure 4.5 Logistic regression at current case

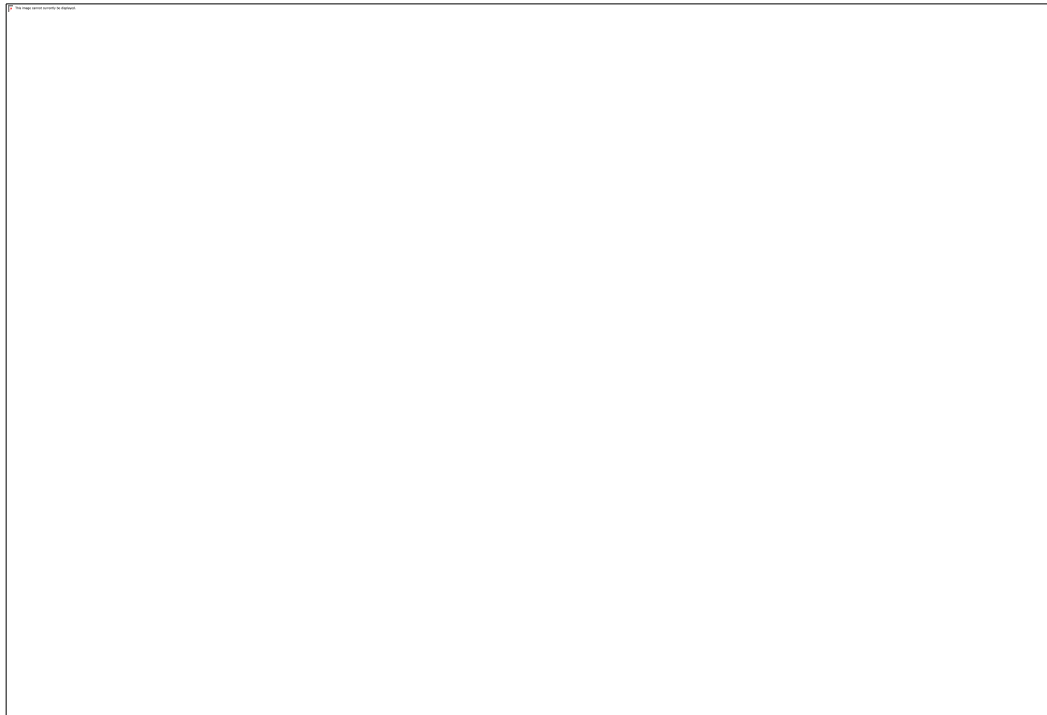


Table 4.10 AUC Logistic regression

Area Under the Curve

Test Result

Variable(s): Predicted

probability.

Area
.639

Figure 4.6 Survival analysis evaluation at current case

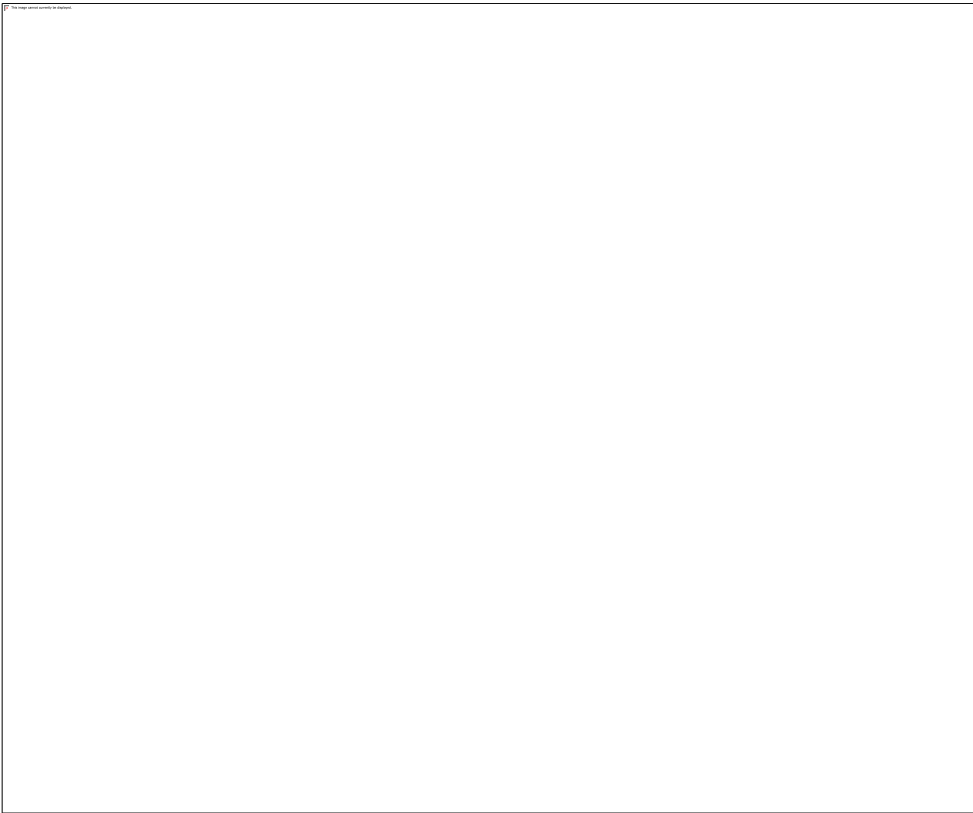


Table 4.11 AUC Survival Analysis

Area Under the Curve

Test Result

Variable(s):Survival function

evaluate at the current case

Area
.717

Comparison of Survival Analysis to Logistic Regression

The comparison of model results obtained from Fig 4.5 and Fig 4.6 respectively, reveals that the survival analysis model outperforms the logistic regression model in terms of predictive accuracy, albeit only slightly. The AUC achieved by the survival analysis model was 0.717, which is higher than the logistic regression model's AUC of 0.639. The survival analysis model's higher AUC value indicates that it can slightly better distinguish between positive and negative outcomes. However, it is essential to note that the logistic regression model also exhibited good predictive performance with an AUC of 0.639. These findings suggest that both the survival analysis and logistic regression models are useful in predicting loan default, with the survival analysis model providing a modest benefit in predictive accuracy. The selection of a model should be based on the specific needs and limitations of the project, as well as further evaluation of the models' performance on the target data.

4.11 Discussion of findings

Both Logistic Regression and Survival Analysis exhibit similar fit according to the ROC analysis. However, the survival model demonstrates superior predictive power because it can better capture the timing of when the client default occurs, compared to the binary outcome predicted by logistic regression. These findings align with previous search Thomal(2014), which

compared the performance of different models and concluded that survival analysis methods outperformed logistic regression. In the Logistic regression model used in this study, several factors were identified as determinants of loan default, loan amount, source of income, home ownership status and loan grade. The analysis, in terms of the logistic regression revealed that income level, loan amount, loan grade, age and purpose of loan had an impact on loan default. Both survival model and logistic regression indicated that age of an applicant had slightly no effect on loan default, which is consistent with the findings of Yegon et al(2014), who conducted a similar study in Kenya and found that personal factors played a significant role in determining the likelihood of default.

4.12 Chapter Conclusion

The chapter consisted of various examinations conducted to evaluate the effectiveness of the models. Such examinations comprised of the Hosmer-Lemeshow goodness of fit test and the diagnostic for correlation. The formulation of models for estimating the probability of default was achieved through logistic regression and survival analysis techniques. A comparative analysis was performed to assess the two models, whereby survival analysis overshadowed logistic regression in terms of competitiveness.

CHAPTER 5: SUMMARY, RECOMMENDATIONS AND CONCLUSIONS

5.0 Introduction

The concluding chapter of the research delves into the findings and addresses each objective accordingly. An overview of the research is also included. Only the findings of this research were used to make recommendations. Furthermore, the conclusion includes a list of areas for future research.

5.1 Summary of findings

The research objective: To develop a model upon lending companies can appraise the creditworthiness of a borrower before granting loan. The model built using logistic regression was used here. The Logistic Regression Model below was found to be handy in predicting the probability that the borrower would default or not.

$$= \frac{\exp(-2.050 - 0.005x_1 - 0.051x_2 - 1.077x_3 + 0.042x_4 + 0.551x_5 + 0.007x_6)}{1 + \exp(-2.050 - 0.005x_1 - 0.051x_2 - 1.077x_3 + 0.042x_4 + 0.551x_5 + 0.007x_6)}$$

Where p is the default probability of the borrower will default. The low probability is considered desirable here which implies that the client has low chance of default and thus less credit risk. The omitted variables were considered insignificant or less significant in the model. They are explanatory variable which were found to have significant effects on loan repayment default are the ones included in the model.

This model can be used to evaluate an applicant's creditworthiness by assessing their default probability. Those with low default probability may be granted a higher loan amount. The applicant's default probability can also be used to calculate the loan interest. The risk associated with a loan portfolio determines the interest charges. Portfolios with higher loan default probability carry higher risk and attract higher interest charges.

The following variables were taken into consideration when creating a model to predict the probability of loan default: age of applicant, loan amount, source of income, debt to income ratio, purpose of the loan and loan term. However, since this research is quantitative, only the quantifiable determinants are presented here. The good will of the applicant and their willingness to pay back the loan are important factors, but they were not included in the model, despite their significance.

This study aimed to create models for predicting loan default in a micro-finance institution in Zimbabwe, using logistic regression and survival analysis. The models take into account changing factors over time and offer a reliable way to forecast the likelihood of default in different periods. The study compared the performance of the two models using real data from a loan portfolio and found that both were competitive, with the logistic regression model ranking similarly well. While more complex models have been explored, their cost and potential lack of improvement in predictive accuracy make survival analysis a straightforward and effective alternative to logistic regression.

The study confirmed that that survival modeling slightly outweighed the logistic regression for estimating the likelihood of delinquency in a single specific period. However, survival analysis has several advantages that make it useful for capital and credit risk management.

Overall, this study provides valuable insights into the use of predictive models for managing credit risk in micro-finance institutions, especially in developing countries such as Zimbabwe. It is important to note that the logistic regression models require the building of different models with different data structures in order to make predictions for two different models.

Survival analysis holds an advantage over logistic regression as it can incorporate the most recent data. In contrast, logistic regression is unable to utilize data from the most recent 24 months if the prediction focuses on the probability of default within that timeframe, as it requires at least 24 months of performance data to observe actual defaults. Stepanova (2014) illustrated that the survival probability can be leveraged to calculate the anticipated profit of a loan. This involves aggregating the present values of each installment, multiplied by the probability of it being paid (the customer's survival probability), and then deducting the loan amount. Utilizing the estimated profit derived from a loan can significantly enhance profit management.

Furthermore, survival analysis offers more comprehensive insights into the expected distribution of the time to default. Despite potential limitations stemming from heavy censoring, the knowledge obtained from the predicted T distribution can still be valuable for profit modeling.

Considering cost-benefit analysis, survival analysis is a pricier alternative while the probability of default remains the central factor for reserve calculations. The precise moment when a loan defaults is less relevant than having knowledge of the chance of default within the next year. The logistic regression model provides the supposed probability of default, whereas incorporating survival analysis modeling requires extra calculations, elevating the implementation costs without a significant value addition. Despite this, in larger scenarios like profitability management, particularly where the micro-finance industry is headed, the supplementary benefits from survival analysis modeling may prove helpful.

5.2 Contribution to the study

This study has expanded the current literature on the use of survival analysis and logistic regression in developing countries like Zimbabwe. The study has also contributed by identifying the determinants of loan default, which can be helpful in credit risk management. Additionally, the research has demonstrated the effectiveness of survival analysis as an alternative to the commonly used binary for analyzing whether a client will default or not.

5.3 Recommendations

The government and lending companies should emphasize the importance of training and educating borrowers, as well as increasing the level of financial inclusion in Zimbabwe. The researcher suggests that training should empower borrowers with practical skills, not just theoretical knowledge, to effectively manage their businesses. These skills should encompass risk and cost minimization, focusing on both short-term profit maximization. This approach could potentially lead to a decrease in non-performing loans, reduced credit risk, increased revenue collection, and ultimately higher profits.

5.4 Credit Appraisal model

Numerous lending institutions have a tendency to minimize risks, offering loans solely to civil servants who have a secure income source. They retrieve the funds directly from the salary

before it reaches the individual's bank account, which significantly minimizes the risk of loans not being repaid. Despite this, the researcher proposes that lending institutions should broaden their lending criteria to include all potential borrowers, not only civil servants, to increase their loan portfolio and maximize returns for their shareholders. However, retrieving loans from non-civil servants can be more challenging; hence continuous monitoring of loan portfolios is essential. To combat this, the researcher recommends utilizing a credit risk evaluation model that can be customized for various types of borrowers, including binary logistic models and survival analysis. These supervised learning models can analyze historical data to identify trends. Zimbabwe has several microfinance institutions adopting different models, including those that offer deposit-taking services and loans geared towards personal development such as Thrive, which only provides loans to women. Since each lending institution's data may be unique, the researcher advises they use their own past data to create a bespoke model for credit assessments.

5.5 Consistent monitoring

The credit risk team should monitor the loan book to avoid potential defaults. Strategies such as debt to property swap, debt restructuring, and legal action can be implemented to reduce losses. Key risk indicator (KRI) reports band trend analysis is useful tools for monitoring loan portfolios

5.6 Conclusions

Two different methods were utilized to generate loan default probability prediction models, survival analysis and logistic regression. The survival model has a greater predictive power when compared to logistic model. Therefore, micro-finance institutions should consider utilizing survival model to decrease the prevalence of non-performing loans. The study also discovered that age of applicant, amount borrowed, source of income; debt to income ratio and loan grade are all determinants of loan default.

5.7 Areas for further research

Political instability can have an impact on loan default rates. The economic factors that influence loan default rates are worth examining. Current credit risk techniques should be evaluated. The management of liquidity risk and credit supply is crucial in financial markets. The impact of the fundamental review of trading book (FRTB) on risk management in investment banks should be

analyzed. The relevance of modern credit score cards in credit risk should be considered. A comparison between impairments under Basel accords and under IFRS9 impairment is necessary.

5.8 Chapter Conclusion

This chapter addresses the research objectives and questions, provides a summary of findings and conclusions, and discusses the method for predicting default probability; the models constructed using logistic regression and survival analysis, and the factors influencing loan default. Additionally, it outlines recommendations for all stakeholders.

REFERENCES

- Aginer, D., Demirguc-Kunt, A. and Zhu, M. (2012) 'How Does Banking Competition Affect Systematic Stability? The World Bank, Development Research Group, Finance and Private Sector Development', Research Working Paper 5981.
- Auronen, L. (2003) Asymmetric Information: Theory and Applications. Paper presented at the Seminar of Strategy and International Business, Helsinki University of Technology, May 21st.
- Ben, M. (2008) The Effect of Dollarization on the Performance of the Zimbabwe Stock Exchange.
- Bercoff, J. J., di Gresia, L. and Grimard, F. (2002) 'Argentinean Banks, Credit Growth and the Tequila Crisis: A Duration Analysis'.
- Bofondi, M. and Gobbi, G. (2003) 'Bad Loans and Entry in Local Credit Markets', Bank of Italy Research Department, Rome.
- Bryman, A. and Bell, E. (2003) Business Research Methods. Oxford: Oxford University.
- Burns, R. and Burns, R. (2008) 'Logistic Regression', in Business Research Methods and Statistics.
- Castro, V. (2013) 'Macroeconomic Determinants of Credit Risk in the Banking System: The Case of GIPSI', Economic Modelling, 31, pp. 672-683.
- Cooper, D.R. and Schindler, P.S. (2008) Business Research Methods. Kent: McGraw-Hill Higher Education.
- Cooper, H. and Hedges, L.V. (eds.) (2003) Handbook of Research Synthesis. New York: Russell Sage Foundation.
- Coyle, B. (2000) Framework for Credit Risk Management. United Kingdom: Chartered Institute of Bankers.
- Das, A. and Ghosh, S. (2007) 'Determinants of Credit Risk in Indian State-owned Banks: An Empirical Investigation', Economic Issues, 12(2), pp. 48-66.
- Dietrich, A. and Wanzenried, G. (2009) What Determines the Profitability of Commercial Banks? New Evidence from Swaziland.
- Garcia-Herrero, A. (2006) 'What Explains Low Profitability of Chinese Banks?', Working Paper No. 30, The American University of Paris.

- Garr, D.K. (2013) 'Determinants of Credit Risk in the Banking Industry of Ghana', *Developing Country Studies*, 3(11).
- Igan, D. and Pinheiro, M. (2011) 'Credit Growth and Bank Soundness', IMF Working Papers WP/11/278, pp. 4-8.
- Jappelli, T. and Pagano, M. (2000) 'Information Sharing in Credit Markets: The European Experience', CSEF Working Paper No. 35, University of Salerno.
- Jensen, M.C. and Meckling, W.H. (1976) 'Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure', *Journal of Financial Economics*, 3(4), pp. 305-360.
- Kwambai, K.D. and Wandera, M. (2013) 'Effects of Credit Information Sharing on Nonperforming Loans: The Case of Kenya Commercial Bank, Kenya', *European Scientific Journal*.
- Kithinji, A.M. (2010) 'Credit Risk Management and Profitability of Commercial Banks in Kenya'.
- Knox, K. (2004) 'A Researcher's Dilemma-Philosophical and Methodological Pluralism', Nottingham Business School, Nottingham Trent University, UK.
- Louzis, D.P., Vouldis, A.T. and Metaxas, V.L. (2011) 'Macroeconomic and Bank-specific Determinants of Nonperforming Loans in Greece: A Comparative Study of Mortgage, Business and Consumer Loan Portfolios', *Journal of Banking & Finance*, 36(4), pp. 1012-1027.
- Mabvure, J.T. and Gwangwava, E. (2012) 'Non Performing Loans in Commercial Banks: A Case of CBZ Bank Limited In Zimbabwe', *Interdisciplinary Journal of Contemporary Research in Business*, 4(7).
- Merrouche, O. and Nier, E. (2010) 'What Caused the Global Financial Crisis-Evidence on the Drivers of Financial Imbalance 1999-2007', IMF Working Paper.
- Mwaura, I.G. (2013) 'The Determinants of Credit Risk in Commercial Banks in Kenya', University of Nairobi.

- Richard, E. (2011) 'Factors That Cause Non-Performing Loans in Commercial Banks in Tanzania and Strategies to Resolve Them', *Journal of Management Policy and Practice*.
- Reserve Bank of Zimbabwe (2006) 'Risk Management Guideline No. 01-2006 BSD'.
- Reserve Bank of Zimbabwe (2000) 'Banking Regulation 2000, 205 of 2000'.
- Reserve Bank of Zimbabwe (2010) 'Technical Guidance on Basel II Implementation in Zimbabwe, Guideline No. 1 2010'.
- Yin, R.K. (2006) *Case Study Research: Design and Methods*. 4th edn. Sage Publications.
- Salas, V. and Saurina, J. (2002) 'Credit Risk in Two Institutional Regimes: Spanish Commercial and Savings Banks', *Journal of Financial Services Research*, 22(3), pp. 203-224.
- Saunders, A. and Cornett, M.M. (2008) *Financial Institutions Management*. 6th edn. McGraw-Hill.
- Saunders, M., Lewis, P. and Thornhill, A. (2009) *Research Methods for Business Students*. 5th edn. Harlow: FT Prentice Hall.
- Stepanova, M. and Thomas, L. (2002) 'Survival Analysis Methods for Personal Loan Data', *Operations Research*, 50(2), pp. 277-289.
- Tamirisa, N.T. and Igan, D.O. (2007) 'Credit Growth and Bank Soundness in Emerging Europe', *IMF Working Paper 2007*, pp. 2-7.
- Tennant, D. and Folawewo, A. (2009) 'Macroeconomic and Market Determinants of Banking Sector Interest Rate Spreads: Empirical Evidence from Low and Middle Income Countries', *Applied Financial Economics*, 19(6), pp. 489-507.
- Thomas, L.C., Oliver, R.W. and Hand, D.J. (1999) 'A Survey of the Issues in Consumer Credit Modeling Research', *Journal of the Operational Research Society*.
- White, B. (2000) *Dissertation Skills for Business Management Students*. Continuum.
- Yin, R.K. (2003) *Case Study Research: Design and Methods*. 3rd edn. Sage.
- Zikmund, W.G. (2003) *Business Research Methods*. 7th edn. South-Western: John Wiley and Sons Inc.

APPENDICES

APPENDIX A: HISTOGRAM OF AGE

```
GET DATA /TYPE=XLSX
  /FILE='C:\Users\User\Documents\D3061300.xlsx'
  /SHEET=name 'Sheet1'
  /CELLRANGE=full
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.
DATASET NAME DataSet1 WINDOW=FRONT.
GRAPH

  /HISTOGRAM(NORMAL)=age.
```

APPENDIX B: DESCRIPTIVE STATISTICS

```
DESCRIPTIVES VARIABLES=age emp_length_int home_ownership_cat income_cat loan_amount pu  
rpose_cat Lc dti  
/STATISTICS=MEAN STDDEV MIN MAX.
```

APPENDIX C: CORRELATION TEST

CORRELATIONS

```
  /VARIABLES=age emp_length_int home_ownership_cat income_cat loan_amount purpose_cat  
Lc dti  
  /PRINT=TWOTAIL NOSIG  
  /MISSING=PAIRWISE.
```

APPENDIX D: TEST FOR MULTI-COLINEARITY (VIF)

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT loan_default
  /METHOD=ENTER age emp_length_int home_ownership_cat income_cat loan_amount purpose_c
at Lc dti.
```

APPENDIX E: LOGISTIC REGRESSION

```
LOGISTIC REGRESSION VARIABLES loan_default
  /METHOD=ENTER age emp_length_int home_ownership_cat income_cat loan_amount purpose_c
at Lc dti
  /SAVE=PRED PGROUP
  /CLASSPLOT
  /PRINT=GOODFIT CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

APPENDIX F: COX REGRESSION

```
COXREG time
  /STATUS=loan_default(1)
  /METHOD=ENTER age emp_length_int home_ownership_cat income_cat loan_amount purpose_c
at Lc dti
  /PLOT SURVIVAL HAZARDS
  /SAVE=SURVIVAL HAZARD XBETA
  /PRINT=CI(95) CORR
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```


APPENDIX G: ROC ANALYSIS

```
ROC PRE_1 BY loan_default (1)
  /PLOT=CURVE(REFERENCE)
  /CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING=EXCLUDE.
```

```
ROC SUR_1 BY loan_default (1)
  /PLOT=CURVE(REFERENCE)
  /CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING=EXCLUDE.
```

LETTER OF CONFIRMATION

KCI MANAGEMENT CONSULTANTS (PVT) LTD

14 GRAND HOTEL

SHURUGWI ZIMBABWE

SUBJECT: CONFIRMATION LETTER FOR REBECCA TAFADZWA MBENGI REGISTRATION NUMBER B202857B TO HAVE ACCESS ON COMPANY DATA FOR ACADEMIC PURPOSES

To whom it may concern

I hereby confirm that REBECCA TAFADZWA MBENGI is authorized to access and use the specified data from KCI Management Consultants for the sole purpose of her academic research. It is understood that the data provided to REBECCA TAFADZWA MBENGI should be used strictly for academic purposes and should not be shared or used for any commercial or non-academic activities. Furthermore, REBECCA TAFADZWA MBENGI is expected to adhere to all confidentiality and data protection policies of KCI Management Consultants while using the provided data.


Should you require any further verification or details regarding this permission, please feel free to contact me.

Sincerely,

MRS T. M Nyanzara

Senior Branch Manager

KCI Management Consultants



KCI MANAGEMENT CONSULTANTS
(PVT) LTD
94 McChisney Avenue
Eastlea, Harare