University of Nairobi

FACULTY OF SCIENCE AND TECHNOLOGY

SCHOOL OF MATHEMATICS

STA 420:PROJECT

LOGISTIC REGRESSION ANALYSIS ON BANK LOAN DEFAULT

Submitted in partial fulfillment for the award in degree of

BACHELOR OF SCIENCE STATISTICS

Declaration

This is to certify that this is a bonafide record of the project presented by the students whose names are given below in partial fulfillment of the requirements of the degree in Bachelor of Science Statistics.

Onyango Tracey Atieno I63/4623/2020

Signature ____________ Date 14/06/2024

Omoro Stacy Nyanami I63/4601/2020

Signature ____________ Date 14/06/2024

Onyango Margaret Tracy I63/4603/2020

Signature ____________ Date 14/06/2024

Akinyi Lucy Marylyn I63/4609/2020

Signature ____________ Date 14/06/2024

Kamau Carolyne Njoki I63/4600/2020

Signature ____________ Date 14/06/2024

In my capacity as a supervisor of the candidates.I certify that this project has my approval for submission.

Signature ____________ Date ____________

Prof.George Muhua(Supervisor)

Dedication

We dedicate this project to our parents, lecturers, relatives and friends.

Acknowledgement

We would like to express our deepest gratitude to God Almighty for His grace throughout the years and during this project.We would wish to thank the University of Nairobi and the Department of Mathematics for the opportunity to showcase what we have learned in our course.We would also like to thank our supervisor Prof . George Muhua, for his invaluable guidance, patience and support.We would also like to thank ourselves for our resilience and teamwork for putting together this project.

List of Abbreviations

CBK:Central Bank of Kenya

NPLs:Non Performing Loans

NACOSTI:National Commission for Science, Technology and Innovation

MFI:Monetary Financial Institutions

ROE:Return on Equity

GDP:Gross Domestic Product

SMEs:Small and Medium Enterprises

OLTV:Originating Loan to Value Ratio

Abstract

An increase in loan defaults have led to financial institutions incurring losses over the years. Addressing the root causes of loan default is imperative for ensuring the sustainability of the banking sector. This research aims to utilize logistic regression to predict probability of loan default with an aim of identifying demographic, socio-economic and loan specific factors associated with loan default.

The primary objective is to model the relationship between various predictor variables and loan default outcome. The specific aims include investigating the role of demographic, socio-economic and loan specific characteristics in predicting default probability and building a logistic a regression model to analyze likelihood of loan default. The study used a quantitative loan dataset sourced from the Bondora public reports page containing 36 variables related to loan characteristics and borrower demographics. Logistic regression analysis was employed to predict probability of loan default based on the dataset used. Bivariate analysis highlighted demographic and socio-economic factors associated with default. Data transformation techniques were applied and the dataset was split into training and testing tests. A logistic regression model was fitted using the training set and backward stepwise regression was employed to select significant variables. Bivariate analysis and correlation analysis were conducted to understand the relationship between loan default likelihood. The analysis showed significant insights into factors influencing loan default. The key findings included: impact of age, gender, marital status, employment status and loan characteristics on default probability. The logistic regression model demonstrated 65% accuracy with a sensitivity of 93% and specificity of 52%.

Recommendations made were exploring advanced machine learning techniques to enhance model accuracy and interpretability, incorporating variables such as co-signers, health and lifestyle factors and political and economic indicators and employing qualitative research methods such as surveys or interviews to improve credit risk assessment and provide deeper insights into the underlying causes of loan default.

Contents

List of Figures

List of Tables

1 Chapter 1 :Introduction

1.1 Background of the study

Loan default is defined as the risk of loss due to the borrower's failure to fulfill the loan contract on time. (Zhang et al., 2023). Loan lending has been playing a significant role in the financial world throughout the years. Although it is quite profitable and beneficial for both the lenders and the borrowers, it carries a great risk, which in the domain of loan lending is referred to as Credit risk. (Aslam et al., 2019). Concurrently, banking authorities are finding it more difficult to correctly assess loan requests and tackle the risks of people defaulting from on loans. The questions that mostly arise are (i) How risky is the borrower? And ii) Given their risk, should we lend him/her? According to Zhao and Zou (2021), prior to the 20th Century money lending processes were highly subjective and borrowers were evaluated by their trustworthiness character wise. The bias in this process naturally led to the creation of credit scores. Most banks assess the credit- worthiness of customers using a credit scoring system. This is a multivariate assessment system that effectively predicts the risk of loan default by weighing the set of defined variables and delivering a single value that shows the risk of default (Mugenda, 2008). As such a low score could be interpreted as a low risk of default that is used to imply that the client is able to service the loan, with a high score indicating a high risk of defaults thus the rejection of the loan request. In Kenya, commercial banks are the key players not only in the financial market but also in spurring the economic growth that has been witnessed in the country in the recent past. The biggest part of these huge profits emanates from the interests charged on loans they advance to their customers. If these loans non-perform, these blue-chip companies will come tumbling down and the entire economy will be threatened.

In response to the evolving economic landscape, banks like Equity Group Holdings PLC have been actively working to make loan repayments more manageable for borrowers. As outlined in Equity Group Holdings PLC's 2022 Integrated Report and Financial Statements, the bank has implemented several initiatives to support businesses facing financial challenges. Some of the initiatives include: loan repayment extensions, introduction of credit guarantees and offering advisory support to foster financial resilience and supporting the long-term success of their clients amidst challenging economic conditions. (EQUITY BANK (KENYA) LIMITED, n.d.) Despite the measures put in place to make loan repayments manageable, a report on Business Daily confirms a rise in loan defaults. By the end of October 2023, the stock of bad loans held by Kenyan Banks grew to 26%. This mounting bad loans are therefore eating into banks' profitability as they increase their loan-loss provisions to cover for defaults. (January 24 2024). Loan default poses a significant risk to both lenders and borrowers, impacting the stability of financial institutions and the broader economy. Hence addressing the root causes of loan defaults and refining credit assessment is imperative for ensuring the sustainability of the banking sector and fostering economic resilience in Kenya. This is why in this paper, the main objective is to use logistic regression to predict loan default in banking.

1.2 Problem Statement

The banking sector is currently facing the acute problem of Non-Performing Loans (NPLs), a sign of ineffective lending practices. The increasing amount of NPLs reduces the bank's profit and the capacity of lending by reducing bankable assets. Some of the credit scoring models used by banks to evaluate the probability of a loan default include traditional credits scoring which assess factors such as income, assets, employment history, and existing debt obligations thus assigning numerical scores to borrowers based on their credit history and financial stability and also credit bureau reports from entities such as the Credit Reference Bureau (CRB) to evaluate the credit history of potential borrowers.

Although traditional credit scoring and credit bureau reports are valuable tools in assessing credit risk, they have limitations in terms of variable selection, adaptability, outcome measurement, data quality, and transparency compared to logistic regression which provides a direct estimate of the probability of a binary outcome, allowing for a more quantitative and probabilistic assessment of credit risk.

1.3 Objectives

The overall objective of this study is to use logistic regression analysis to predict loan default in banking. The specific objectives include:

1. To identify the demographic and socio-economic factors associated with bank loan default in Kenya.

2. To investigate the role of loan-specific characteristics including loan amount, interest rate and loan purpose in predicting default probability.

3.To fit a logistic regression model to analyze the likelihood of loan default based on the identified factors.

1.4 Justification

This study is essential for addressing a pressing concern within Kenya's banking sector: the escalating rate of loan defaults. With the potential to destabilize financial institutions and hinder economic growth, understanding the determinants of loan default behavior is paramount. By employing logistic regression analysis, this study aims to fill a crucial gap in current credit risk assessment practices, which often overlook demographic and socio-economic factors. Through a deeper understanding of these factors and their impact on default probability, this study can contribute to the development of more accurate predictive models, enabling banks to proactively identify at-risk borrowers and implement targeted risk management strategies. Furthermore, the findings of this study can inform policymakers and regulatory authorities in crafting more effective frameworks for monitoring and managing credit risk, thereby promoting financial stability and consumer protection. Ultimately, by addressing the root causes of loan defaults and refining credit risk assessment techniques, this study has the potential to enhance the resilience of Kenya's banking sector and foster sustainable economic development.

1.5 Significance of the study

The study on predicting loan default in banking through logistic regression holds immense significance within Kenya's economy. Loan default poses a substantial risk to financial institutions and the broader economy, as evidenced by the escalating rate of non-performing loans (NPLs) observed in recent years. The banking sector's gross NPLs have surged, threatening financial performance and reducing lending capacity, which undermines investor and depositor confidence. Traditional credit scoring methods and credit bureau reports, while valuable, have limitations in effectively assessing credit risk.

By employing logistic regression analysis, this study seeks to bridge this gap by identifying demographic, socio-economic, and loan-specific factors associated with loan default in Kenya. Through the development of a robust predictive model, banks can proactively identify at-risk borrowers and implement targeted risk management strategies. Moreover, the study's findings can inform policymakers and regulatory authorities in crafting more effective frameworks for monitoring and managing credit risk, thereby promoting financial stability and consumer protection. Ultimately, by addressing the root causes of loan defaults and refining credit risk assessment techniques, this study aims to enhance the resilience of Kenya's banking sector and foster sustainable economic development.

1.6 Limitations of the Study

The study relied on secondary data from Bondora rather than primary data. Due to time constraints we were unable to get the necessary approvals (e.g NACOSTI certificate) to obtain the data from local financial institutions. Some variables in the dataset had missing values or were made anonymous which led to less accurate analysis. An example was the undefined category for gender and the "Other" category for occupation area. Due to the specificity of the dataset to Bondora, the demographic and socio economic factors of loan borrowers in other regions may not be fully represented hence limiting the generalizability of the results to diverse banking environments.

2 Chapter 2:Literature Review

2.1 Introduction

Bank loan defaults, characterized by the inability of the borrower to meet their loan repayment obligations, have significant repercussions on the banks and the financial system's stability. Understanding the causes, consequences, and mitigation strategies of bank defaults is essential for policymakers, regulators, investors, and researchers alike.

This literature review delves into the theoretical underpinnings of bank loan defaults, exploring how different frameworks illuminate the complexities of borrower-bank dynamics and the potential causes of loan failure. By analyzing the theories from a wide range of academic studies, we shed light on the key drivers of bank defaults, the mechanisms through which they unfold, and the methods employed to prevent and manage them

2.2 Behavioural Finance Theory

Traditional finance assumes borrowers act perfectly logically and have the information they need when applying for a bank loan.Shefrin(2002), challenges this notion, that the borrowers' emotions and behavioral biases can significantly influence loan default. He highlights how several biases such as overconfidence, optimism, and trend-chasing can negatively impact borrowers.

The results lead borrowers to underestimate the true cost of borrowing, take on more loans than they can manage, and delay payments when facing financial difficulties. Banks should develop strategies such as financial literacy programs and design loan products with realistic payment plans for borrowers.

2.3 Agency Theory

It mainly considered the relationship between an agent and a principal. According to (Ekanayake, S. 2004) in his journal, agency theory mainly considers the relationship between an agent and a principal. Agents were entrusted with decision-making authority or resources by principals An example is in banks where the agents are the bankers while the principals are the owners of the banks.

Agency theory attempts to resolve two problems relating to the agency problem. The first is the monitoring problem that arises because the principal cannot verify whether the agent has behaved appropriately. Agents can exert little effort without being observed causing a moral hazard problem. (Holstrom, 1979). The second is the problem of risk sharing (particularly in the case of outcome-based controls) that arises when the principal and the agent have different attitudes toward risk (Eisenhardt, 1989; Jensen and Meckling, 1976). Informational asymmetries between borrowers and lenders can aggravate agency problems, leading to adverse selection and moral hazard issues where agents have incentives to engage in behaviors that increase their utility or welfare at the expense of the principals. Adverse selection is where high-risk borrowers might actively seek loans, leading banks to unknowingly lend to riskier clients (Mishkin, 2007) while moral hazard is when the borrower might have less incentive to manage funds well, increasing default risk. He proposes that principals employ various mechanisms, such as monitoring through information systems such as management control systems, incentives, and contractual arrangements. Bank loan defaults are a complex issue with roots in various theoretical frameworks. By understanding the interplay of these theories, banks can develop more robust lending practices, improve risk assessment, and ultimately reduce defaults.

2.4 Review on Loan Characteristics

Loan terms, encompassing elements such as interest rates, loan amounts, loan-to-value ratio, repayment schedule and loan purpose play a pivotal role in shaping borrowers' repayment behaviors and, consequently, the overall credit risk profile of lending portfolios. Loan-to-value ratio (LTV), the measure of the borrowed amount relative to the collateral's value, is crucial in assessing default risk. The European Central Bank (2018) conducted a study, "The impact of lending standards on default rates of residential real estate loans" to assess the influence of lending standards (loan-value ratio and loan-income ratios at origination, maturity of the original loan as well as the borrower's status of employment) on default rates. The loan data used was from multiple euro area countries. The data analysis was conducted through descriptive analysis and static probit regression. The results obtained indicated a high correlation between loan-to-value ratio at the beginning (OLTV) and increased loan default probability. The main regression results show that OLTV is significant with a p-value of (0.003) at 1% significance level with a positive coefficient of 0.0002. Hence indicating that when the OLTV ratio increases by 1%, the chance of defaulting goes up 0.02%. Ultimately, if a borrower borrows more money relative to the value of the collateral, the probability of defaulting the loan repayment slightly increases. In relation to this, other findings by Ranisavljević (2016) on "Realistic Evaluation of the Ratio-The Key to Minimizing the Credit Risk" suggested that commercial banks should establish initial LTV ratios for various long-term loan products to prevent their increase over time thus mitigating loan default by taking into account all the factors that cause the ratio's increase, hence focusing on reducing the loan defaults other than making loans more attractive and accessible.

Itoo et al (2013) conducted a study on the factors influencing mortgage default. The researchers analyzed data from Jammu and Kashmir banks from secondary but not published data. The data encompassed defaulters from April 2011-March 2012. A sample size of 115 was chosen through systematic random sampling. The study period spread from September 2011 to August 2012. Several statistical tests were employed using SPSS 17.0 to analyze the results, i.e, chi-square, correlation, regression, ANOVA, paired 't' test and logistic regression. The results indicate that the loan rate (r=0.424, p-value < 0.01) and loan amount (r=0.847, p-value < 0.01) have a positive relationship with the outstanding balance at 99% level of confidence while amount repaid (r=0.036, p-value > 0.05), purpose of the loan (r=0.048, p-value > 0.05) and loan type (r=-0.155, p-value >0.05) have a weak relationship with the outstanding loan balance at 95% level of confidence. These findings suggest that factors such as loan rate, amount repaid, purpose of the loan and loan amount influence mortgage default. The researchers employed various statistical techniques hence enabling thorough examination of the relationships between different variables and mortgage default. Additionally, Mabeya (2019) sought to find out the determinants of loan repayment defaults among Small and Medium Enterprises (SMEs) in Kisumu Central Sub-County, Kenya. He focused on various aspects of loan repayment default such as ownership (sole proprietorship, partnership, registered company), borrower (age, gender, income status, credit use), loan (loan size, value of collateral, repayment period) and lender characteristics (time between initial application and payment, financing cost, capacity to approach business data). A descriptive cross sectional study design was used whereby the researcher administered closed ended questionnaires to 600 SMEs who had benefitted from loan. Stratified random sampling was used to sample the SMEs. The study found out that loan characteristics influenced loan default moderately (Mean = 3.7594) while loan amount and collateral value influence loan default to a higher extent (Mean = 4 and above), most of the respondents agreed that ownership characteristics determine loan repayment default (Mean = 3.5456), borrower's characteristics moderately affect moderately affect loan default (Mean = 3.7778) with their education level, existence of other businesses and attitude influencing loan defaults to a greater extent (Mean = 4 and above) and lender characteristics affect loan default (Mean = 3.7115). While Mabeya's study sheds light on important factors influencing loan repayment defaults among SMEs in Kisumu Central Sub-County, there are areas where the analysis, and interpretation could be strengthened to enhance validity of the findings. The interpretation of "moderately affect" and "influence to a higher extent" lacks precision and could benefit from statistical tests. Further research using more robust methodologies and analytical techniques would help to deepen our understanding of the determinants of loan default in this context.

2.5 Review on Borrower Characteristics

Examining the influence of borrower characteristics on loan default in banks has provided valuable insights into the factors that contribute to credit risk. Overall, evidence suggests that borrower characteristics play a crucial role in determining loan default probabilities. Banks and financial institutions utilize this knowledge to develop robust credit risk assessment models and tailor lending practices to mitigate default risk effectively. However, it's essential to recognize that default prediction is a complex and multifaceted process influenced by various interrelated factors, requiring ongoing research and refinement of analytical techniques.

In addition(Vandell and Thibodeau 1985) study consistently found that borrowers with lower credit scores or poor credit histories are more likely to default on loans. A credit score is often regarded as one of the most reliable predictors of default risk, with lower scores indicating higher risk.

Income and employment stability is an important factors to consider when evaluating the likelihood of default. According to (Lin et al., 2011) borrowers with unstable income sources or frequent job changes are more likely to default on loans. Employment stability and steady income are considered essential factors in assessing a borrower's ability to repay debts. A high debt-to-income ratio, indicating that a significant portion of the borrower's income is already committed to debt repayment, is associated with an increased likelihood of default. Borrowers with a higher debt burden may struggle to meet additional repayment obligations, increasing default risk.

Furthermore, loan purpose positively correlates with loan default. The study by (Netzer et al., 2019) suggests that the purpose of the loan can influence default rates. For example, loans taken for investment purposes might have different default rates compared to those taken for consumption or emergencies. Business loans, particularly for startups or small businesses, often carry higher default risks compared to personal loans due to the inherent uncertainties in business ventures. Loan-specific factors, such as loan amount, interest rate, and term length, can influence default probabilities. Higher loan amounts relative to borrower income, as well as unfavorable loan terms, may increase default risk. Loan-specific factors, such as loan amount, interest rate, and term length, can influence default probabilities. Higher loan amounts relative to borrower income, as well as unfavorable loan terms, may increase default risk.

Demographic factors such as Age are significant in predicting loan default Study by(Lin et al., 2011; and Nigmonov et al., 2022) explains the correlations. between borrower age and default rates, with younger borrowers or those in certain age brackets exhibiting higher default risks. Demographic factors such as marital status, education level, and geographic location have also been investigated, although findings vary across studies.

The economic state of a country should be taken into account when looking into default. A study in the US by (Nigmonov et al., 2022) stated that economic factors, such as unemployment rates, GDP growth, and interest rate fluctuations, can significantly impact borrower default rates. During economic downturns or periods of financial instability, default rates tend to rise as borrowers face increased financial stress.

2.6 Review on Lender/ Institutional Characteristics

While borrower characteristics undoubtedly play a role, lender/institutional characteristics contribute to default rates. This review explores how various aspects of a bank's practices influence loan repayment.

According to the scholarly works of Abid, Ouertani, and Zouari-Ghorbel (2014), we determine that management quality plays a pivotal role in determining loan defaults in the Tunisian banking sector. The study investigated Macroeconomic and Bank-Specific Determinants of Household Non-Performing Loans in Tunisia. The research employed panel data analysis estimated from 2003 to 2012 on about 16 Tunisian banks. Results indicated that apart from macroeconomic variables such as GDP, interest rates, and inflation rates, non-performing loans could also be explained by bad management quality. Findings showed that the coefficient of the inefficiency index was positive and statistically significant. This provided support for the 'bad management' hypothesis which implied that banks with poor management practices are more likely to experience higher non-performing loans due to their inability to effectively assess borrower creditworthiness and enforce repayment returns. This further implied that Tunisian banks exhibited tendencies to grant credits with poor quality and also lacked sophisticated evaluation methods to detect insolvent creditors early. The study also explored the relationship between management efficiency, performance indicators, and loan defaults. The significant and negative association between return on equity (ROE) and NPLs for consumer loans provides evidence indicating that the quality of management influences the efficiency of credit granting procedures, particularly those based on quantitative modeling techniques. Moreover, both performance and inefficiency serve as indicators of management quality, with lagged inefficiency positively related to NPLs. This suggests that banks with lower management efficiency may experience higher levels of loan defaults. Overall, these findings underscore the importance of effective management practices in mitigating default risk and enhancing financial performance.

Furthermore, the study provided insights into the risk behavior of banks, highlighting the influence of capital adequacy on loan defaults. The negative and significant explanatory power of the solvency ratio for NPLs supported the 'moral hazard' hypothesis which suggested that banks with low levels of capitalization could be more prone to increased non-performing loans. This is because thinly capitalized banks may engage in excessive risk-taking behavior which consequently result in higher volumes of NPLs due to their riskier lending practices. This finding underscores the importance of capital adequacy in mitigating default risk and maintaining financial stability. The empirical evidence from the Tunisian banking sector highlights the significant role of lender characteristics, particularly management quality and risk behavior, in influencing loan defaults. Effective risk management practices, prudent lending decisions, and capital adequacy are crucial for mitigating default risk and maintaining financial stability. Louzis, Vouldis, and Metaxas (2012) identify performance and inefficiency measures as probable leading indicators for future problem loans. This suggests that regulatory authorities should focus on managerial performance to detect banks with potential NPL increases. The study also mentions that emphasis should be placed on risk management systems and procedures followed by banks to possibly prevent future financial instability.

Nawai and Shariff (2012) examined factors affecting repayment performance in microfinance programs in Malaysia and highlighted the importance of loan monitoring procedures in connection with loan default behaviors. The study employed a survey to collect data on loan repayment performance in microfinance programs in Malaysia. The sample size comprised 309 respondents who were clients of TEKUN Nasional in Peninsular Malaysia. The research revealed ten factors affecting borrower repayment: borrower demographics (age, gender), business characteristics (experience, formality, total monthly sales), loan aspects (distance to the lender's office, loan approval period, loan monitoring practices), and total household income. According to Addae-Korankye (2014), among the significant factors that can increase borrowing transaction costs and also affect repayment performance is loan disbursement lag. Timely loan monitoring and collection with strong follow-up mechanisms are crucial for identifying potential problems early, preventing borrower complacency, facilitating communication, and ensuring the effectiveness of collection strategies. This ultimately helps to minimize loan defaults. The study also similarly identified a lack of monitoring or improper monitoring among some of the major factors of default/delinquency in microfinance institutions in Ghana. Lags between loan disbursement and collection efforts, coupled with weak follow-up mechanisms, can contribute to defaults. Recommendations given include regular review of credit risk techniques used and expanding loan monitoring framework for effective credit portfolio assessment.

A study investigating institutional factors affecting loan default in Kenyan MFIs (Eric Maina, 2014) provides a valuable starting point for understanding the influence of credit policy. The specific credit policy elements explored in this study included the condition of loan facilities (poor or good), strict or lax repayment policies, and well-formulated policies. The target population was 59 MFIs under the Association of Microfinance Institutions of Kenya. A descriptive survey design was used to carry out a census of 59 microfinance institutions in Kenya. A structured questionnaire was administered to the MFI loan officers for response. The model used was Multiple regression to establish relationships between loan default and loan recovery procedures, credit policies, and initial loan appraisal in MFIs in Kenya. 45 out of 48 questionnaires were adequately responded to and considered for analysis. Results show that 31% (the highest percentage of the categories) of the respondents indicated that poor loan facilities have the greatest contribution to loan default rates. When asked about strict repayment policies, the majority (33%) indicated that it has the least contribution, and under well-formulated policies, 31% marked it off as having the least contribution to loan default behavior. Additionally, the majority of the respondents (50%) classified poor record-keeping policies as having the greatest contribution to loan defaulting while 44% agreed that the overall credit policies affect loan defaulting. The findings revealed that credit policy had a significant impact on loan default behavior. Kariuki (2014) also recommended the installation of a vibrant credit process encompassing issues of accurate customer selection, thorough credit analysis, authentic sanctioning process, adequate monitoring, and clear recovery strategies for non-performing loans.

2.7 Review on Ecomomic Conditions

A substantial amount of research suggests that the state of the economy significantly influences the levels of loan losses experienced by banks (De Bock and Demyanets, 2012).The commonly utilized factors for assessing the fluctuations in Non-Performing Loans (NPLs) are outlined below:

A significant portion of research relies on the unemployment rate as an indicator of a nation's economic condition and to explain the decline in loan quality within banks.

Theoretical predictions suggest that fluctuations in unemployment are a potential cause of credit risk (Abras and Rocha, 2020). (Yuksel, 2017) claims that there exists a linear relationship between non-performing loans and the unemployment rate. Unemployment continues to be a pressing issue in many developing nations, including Kenya. As per the Central Bank of Kenya (CBK) data, the banking sector disbursed personal and household loans totaling Sh843.5 billion by the end of December 2021, with Sh70.1 billion categorized as non-performing. This underscores the importance of comprehending the impact of unemployment on credit risk, particularly during periods of record-high unemployment in Kenya. Inflation is a measure of how fast prices of goods and services are rising. It is a key macroeconomic determinant of NPLs.Higher inflation can either have a positive or negative effect on NPLs (Beaton et al., 2016) .The effect of inflation can be uncertain. While higher inflation may ease debt servicing by decreasing the real value of outstanding loans, it can also diminish borrowers' real income if wages remain unchanged.

(Klein, 2013) presents evidence indicating that during periods of inflation, borrowers encounter greater difficulty in repaying their debts, particularly concerning loans with variable interest rates. Jabbouri and Nailli (2019) supported these results. On the contrary, an opposing body of literature indicates a negative correlation between inflation and NPLs (Makri et al., 2014).These studies argue that elevated inflation diminishes the value of outstanding debts, thereby enhancing the repayment capacity of households and firms (Nkusu, 2011).

Researchers agree that under favorable economic conditions, both households and firms tend to be more inclined to fulfill their financial commitments. Due to the inverse relationship between NPLs and economic cycles, wherein NPLs are typically lower during periods of economic expansion and higher during downturns, there exists a negative correlation between GDP growth rates and elevated levels of NPLs (P.K Ozili, 2015).This negative correlation between GDP growth and NPLs stems from the limited revenue stream of borrowers during challenged times leading to a rise in NPLs (Vouldis and Louzis, 2017)

2.8 Summary of the Literature

A few studies have been conducted in an attempt to resolve the suppressing issue of loan default, especially in the banking sector. Most studies indicate that loan default is greatly influenced by the demographic information of the borrowers. Overall from the studies done, we can see a trend from the findings that loan default in Kenyan banks covers various aspects, including causes, consequences, and mitigation strategies. Common causes identified include economic downturns, poor credit assessment, inadequate collateral, and furthermore demographic information of borrowers such as unemployment and age. Consequences of loan default range from financial losses for banks to broader economic impacts such as decreased lending and economic slowdown. Mitigation strategies often involve improving credit risk assessment, implementing effective loan monitoring systems, and promoting financial literacy among borrowers. Additionally, studies explore the role of regulatory frameworks and government policies in addressing loan default challenges in the Kenyan banking sector.

3 Chapter 3:Methodology

3.1 Introduction

This section presents the study methodology used in order to achieve the objectives of the study. This chapter provides the data source, research design and the statistical model used for analysis development process.

3.2 Research Design

This Research Project considers the problem of selecting the relevant predictor variables. Based on the variables identified during the variable selection process, the goal is to predict if a specific borrower is likely to default on their loan. The design involves sourcing data, data processing, data analysis and model fitting by applying binary logistic regression.

3.3 Data Design

This study utilized secondary data that was obtained from Bondora, a peer-to-peer lending platform headquartered in Estonia through their public reports page. The data is accessible through the following link: https://www.bondora.com/en/public-reports. The dataset included loan data from the year 2009 to 2024. For the purpose of this study, the dataset for the years 2015 to 2016 is considered.

3.4 Research Hypothesis

Hypothesis 1

H_0: Demographic and socio-economic factors are not significantly associated with bank loan default

H_1: Demographic and socio-economic factors are significantly associated with bank loan default.

Hypothesis 2

H_0:Loan-specific characteristics do not significantly affect the probability of loan default.

H_1: Loan-specific characteristics significantly affect the probability of loan default.

Hypothesis 3

H_0:A logistic regression model does not accurately predict the likelihood of loan default.

H_1:A logistic regression model accurately predicts the likelihood of loan default.

3.5 Statistical Model

3.5.1 Binary Logistic Regression Model

Binary logistic regression is a statistical method used to model the relationship between a binary response variable and one or more predictor variables. The model combines one or more predictor variables to estimate the probability of occurrence of an outcome of interest. It also gives information on how much an increment in a given predictor variable affects the odds of the response variable. The model can be either a simple model or a multivariable model. The simple model is used to model the relationship between one response variable and one predictor variable while the multivariable model is used to model the relationship between one response variable and two or more predictor variables. Since the response variable is non-numeric, a dummy variable is created which is of the form:

$$Defaultstatus = \begin{cases} 0 & if \quad borrower \quad did \quad not \quad default \quad on \quad the \quad loan \\ 1 & if \quad borrower \quad defaulted \quad on \quad the \quad loan \end{cases}$$

The response variable is a random variable that follows the Bernoulli distribution.

3.5.2 The Bernoulli Distribution

The Bernoulli distribution is a discrete probability distribution used to model experiments that results in either success or failure. For a binary response variable x_i assuming only two values which for purposes of this study were $x_i = 1$ if the ith borrower defaulted and $x_i = 0$ if the borrower paid is a realization of a random variable X_i that can take the values 1 and 0 with probabilities p and 1-p respectively. The distribution is Bernoulli distribution with parameters p which can be written as:

$$f(x) = p^x(1-p)^{1-x} \qquad\qquad x = 0,1$$

for $0 < p < 1$.

3.5.3 The Logit Transformation

The logit transformation is a mathematical function commonly used in statistics, particularly in the context of logistic regression. It transforms the probability of a binary outcome (e.g., success or failure) into a continuous variable that ranges from negative to positive infinity.

The logit transformation is defined as:

$$logit(p) = log\left(\frac{p}{1-p}\right)$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Where:

p=the probability of the event of interest.

1-p= probability of the complementary event.

log= natural logarithm

The resulting logit transformation can take any real value which makes it suitable for regression analysis, where the relationship between predictor variables and the logit of the outcome variable is modeled linearly. In logistic regression, the logit transformation is used to model the relationship between one or more predictor variables and the log odds of the binary outcome. The coefficients estimated in logistic regression represent the change in the log odds of the outcome associated with a one unit change in the predictor variable, holding other variables constant. The inverse of the logit transformation is the logistic function which maps the log odds back to probabilities:

$$logistic(x) = \left(\frac{e^x}{1+e^x} \right)$$

This logistic function allows us to convert the logit-transformed values back into probabilities which are bounded between 0 and 1.

3.5.4 Binary logistic Regression model

The logistic regression model with only one predictor X's is of the form;

$$logit(p) = ln(oddsofanevent) = log\left(\frac{p}{1-p} \right)$$
$$= \beta_0 + \beta_1 x$$

Where; β_0 and β_1 ...β_k are the regression coefficients. To determine the likelihood of the event of interest we utilize the odds ratio formula:

$$odds = \left(\frac{p}{1-p} \right)$$
$$= e^{\beta_0+\beta_1 x}$$

Here, p represents the probability of the event occurring. Alternatively, the probability p can be expressed as:

$$p = \left(\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \right)$$

This equation models the log odds of the event of interest as a linear combination of the predictor variables. The coefficients β_0, β_1,... β_k, are estimated from the data using maximum likelihood estimation, and they represent the change in the log odds of the outcome variable for a one-unit change in the corresponding predictor variable, holding all other variables constant.

3.6 Assumptions

- The response variable is binary.
- The observations are independent.
- There is no multicollinearity among the explanatory variables.
- There is a linear relationship between explanatory variables and the logit of the response variables.
- The sample size is sufficiently large.

3.7 Interpretation of the model coefficients

To accurately interpret the coefficients in our logistic regression model, we need a clear understanding of the relationship between our predictor variables and the probability of default. These variables can either be categorical or continuous and their interpretation varies.

The coefficients for the categorical variables show the change in the log-odds of default compared to a reference category. They show the direction and strength of the effect each category has on the likelihood of default. For example, suppose the coefficient for the ”Unemployed” in the ‘Employment status‘ predictor is positive. In that case, it indicates ’unemployed’ increases the log-odds of the outcome occurring when compared to the reference category (the employed group).

The coefficients for continuous predictor variables indicate the change in log-odds or odds ratio for a one-unit increase in the predictor variable. A positive coefficient indicates an increase in the log-odds or odds ratio as the predictor increases by one unit, while a negative coefficient suggests a decrease. An example, a positive coefficient for ’loan amount’ suggests that the log-odds of default increase with each additional unit of money borrowed which means a higher loan amount leads to a higher risk of default.

3.8 Variable selection

Variable selection in logistic regression refers to the process of choosing which predictor variables to include in the model. Stepwise regression is an iterative process for building a regression model, where independent variables are selected or removed one at a time based on their statistical significance. This method involves a step-by-step approach of adding or eliminating potential predictors from the model, followed by testing for their statistical significance in each iteration.

Backward stepwise selection is a variable selection method whereby a model is fit with all predictor variables and then the least significant variables are iteratively removed until the desired model is achieved. The significance of variables is assessed using criterion such as p- values, AIC (Alkaline Information Criterion) or BIC (Bayesian Information Criterion)

3.9 Model Evaluation

3.9.1 Significance of predictors(Wald's Test)

Wald Test is a test that helps to determine whether individual predictor variables in the model significantly contribute to explaining likelihood of loan default. In logistic regression, this technique validates the significance of predictors in the model. The null hypothesis states that:

H_0=β_i=0

vs

H_1=$\beta_i \neq$0

$$W = \left(\frac{\beta_i}{s.e(\beta_i)}^2 \right)$$

Where: β_i is the estimated coefficient of the predictor variable that is tested.

$s.e(\beta_i)$(standard error of β_i) is the standard error of the estimated coefficient.

The predictor variables are said to be statistically significant if the null hypothesis is rejected.Reject the null if the p-value is less than alpha or when the test statistic greater than critical value(chi-square statistic).

3.9.2 Significance of model(Goodness of Fit)

The likelihood ratio test assesses the goodness of fit of two nested models.The test statistic is:

$$LRT = 2\left(\frac{log-likelihood(fullmodel)}{log-likelihood(nestedmodel)}\right)$$

which follows under the hypothesis that a chi-square distribution equals to the difference in the number of the parameters between the nested model and the full model.If the p-value is less than the level of significance reject the null(use of nested model) in favor of the full model

3.9.3 Confusion Matrix

A confusion matrix is a performance evaluation tool representing the accuracy of a classification model. It displays the number of true positives, true negatives, false positives, and false negatives. This matrix aids in analyzing model performance, identifying mis-classifications, and improving predictive accuracy. For a binary classification problem, we would have a 2 by 2 matrix as shown below, with 4 values:

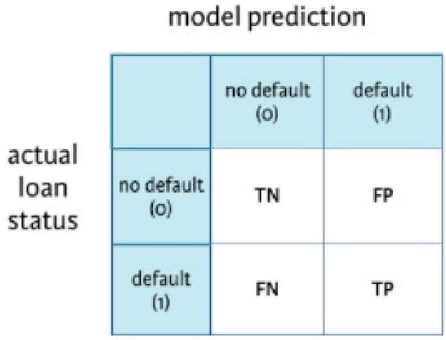

Figure 1: Confusion Matrix

The rows represent the actual values of the target variable. The columns represent the predicted values of the target variable.

In a confusion matrix: True Positive(TP)=the number of instances correctly predicted as positive.

False Positive(FP)=the number of instances incorrectly predicted as positive(actually negative)

True Negative(TN)=the number of instances correctly predicted as negative.

False Negative(FN)=the number of instances incorrectly predicted as negative (actually positive).

We can derive numerous evaluation metrics from the confusion matrix:

Classification accuracy- it refers to the overall correctness of predictions made by a classification model

$$Accuracy = \left(\frac{TP+TN}{TP+FP+TN+FN}\right)$$

Sensitivity-also known as the true positive rate, measures the proportion of actual positives that are correctly identified by the model.

$$Sensitivity = \left(\frac{TP}{TP + FN} \right)$$

Specificity-also known as the true negative rate, measures the proportion of actual negatives that are correctly identified by the model.

$$Specificity = \left(\frac{TN}{TN + FP} \right)$$

4 Chapter 4:Data Analysis

4.1 Introduction

This chapter presents the results, the analysis and the discussion of the dataset used for the purpose of this study.

4.2 Dataset Description

This research utilized loan dataset, which can be found on the Bondora public reports page. This dataset contains 36 variables including loan characteristics and anonymous information about the borrowers, which includes demographic information such as age, gender, country, marital status, education level, employment status and number of dependents. The main goal of this project is to predict whether a borrower will default on a loan or not. This dataset was chosen for its relevance to the study's objectives offering accessible and comprehensive information on bank loan default. We aim to use the data to come up with a predictive model, to identify the demographic and socio-economic factors associated with bank loan default, to investigate the role of loan-specific characteristics including loan amount, interest rate and repayment period in predicting default probability. The following is a brief description of the variables contained in the dataset; LoanNumber which is a unique ID given to all loan applicants. Gender with 0 representing male, 1 representing female and 2 representing undefined. MaritalStatus with 1 representing married, 2 representing cohabitants, 3 representing single, 4 representing divorced and 5 representing widow. EmploymentStatus with 1 indicating unemployed, 2 indicating partially employed, 3 indicating fully employed, 4 indicating self-employed, 5 indicating entrepreneur, 6 indicating retiree. Education with 1 representing primary education, 2 representing basic education, 3 representing vocational education, 4 representing secondary education, 5 representing higher education. Additionally, the variable Amount in the dataset represents the amount the borrower received on the primary market. Interest represents the maximum interest rate accepted in the loan application. LoanDuration represents the current loan duration in months.

4.3 Demographic and Socio-Economic Characteristics

The data contains the personal attributes of the customers including age, gender, marital status, education, employment status, number of dependents and home ownership type. The following plots represent the overall demographic characteristics in the dataset.

4.3.1 Age

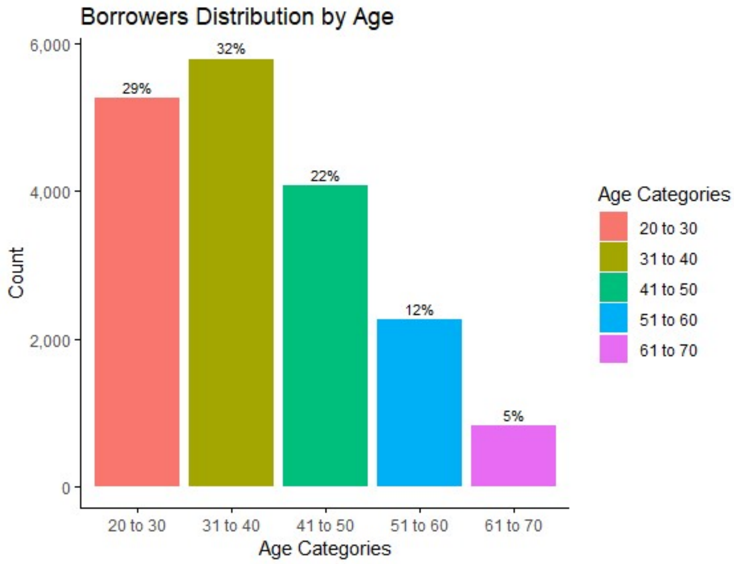

Figure 2: Bar graph of Age Categories

The age group 31-40 recorded the highest proportion of borrowers (31.8%) while the age group 61-70 recorded the lowest proportion of borrowers (4.5%)

4.3.2 Gender

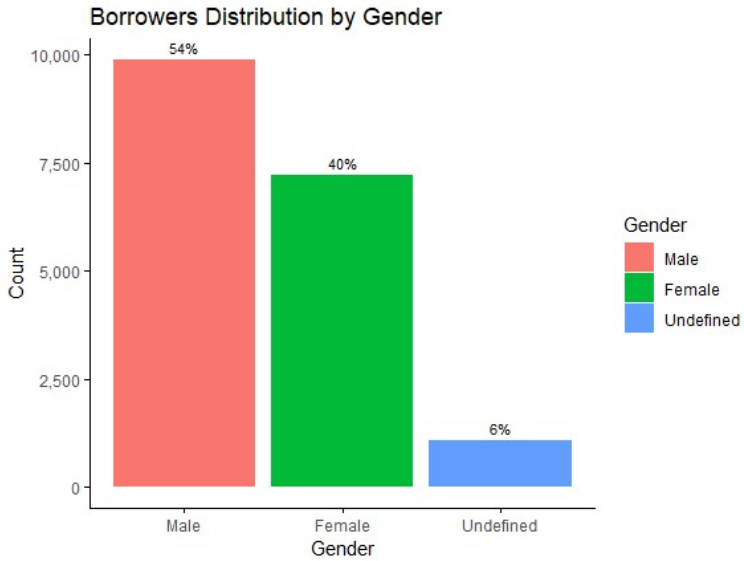

Figure 3: Bar Graph of Gender

Out of the total number of borrowers majority were male (54%), 40% were female and 6% were undefined.

4.3.3 Education

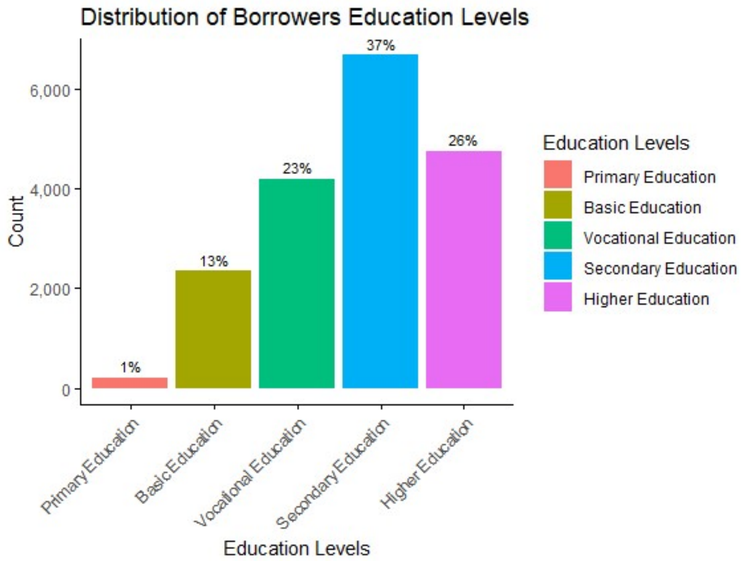

Figure 4: Bar Graph of Education

Most of the borrowers attained up to secondary level of education (36.7%) while the least proportion of borrowers attained up to primary level of education (1.2%)

4.3.4 Employment

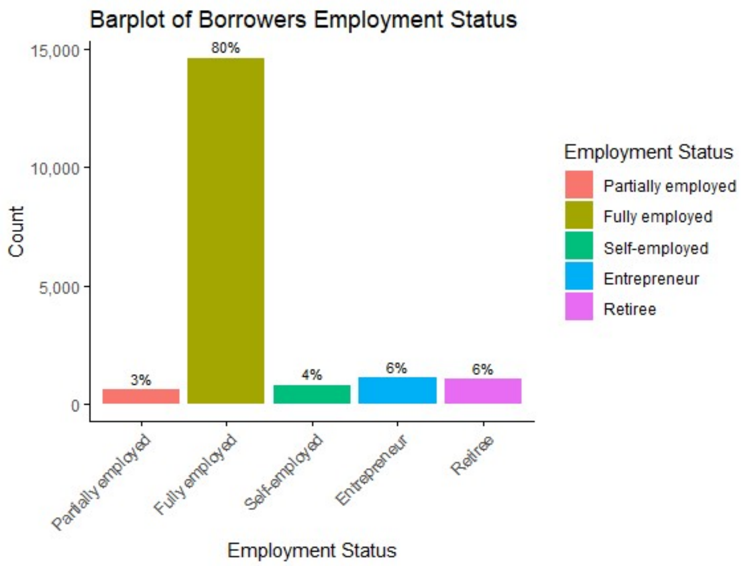

Figure 5: Bar Graph of Employment

Majority of the borrowers were fully employed (80.18%) while the rest were entrepreneurs (6.16%), retirees (5.8%), self-employed (4.42%) and partially employed individuals (3.44%)

4.3.5 Marital Status

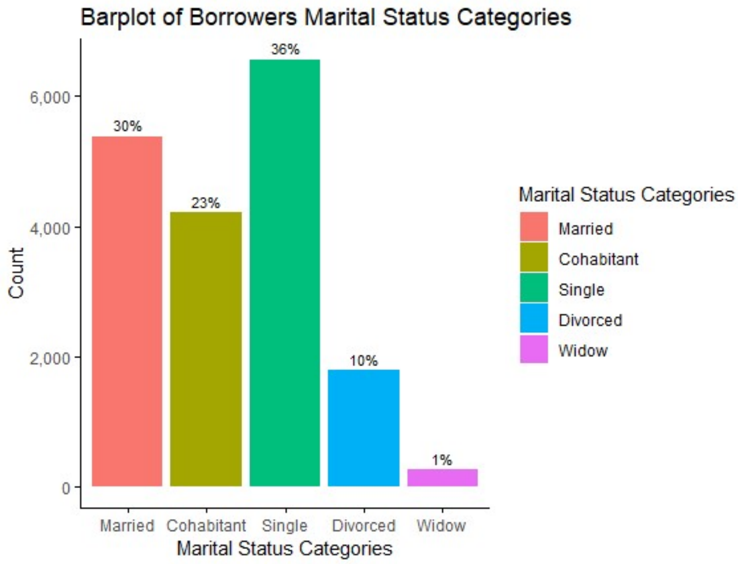

Figure 6: Bar Graph of Marital Status

Majority of the borrowers were single (36%) while the rest were married (29.6%), cohabitant (23.1%), divorced (9.8%) and widowed (1.5%)

4.3.6 Loan Characteristics

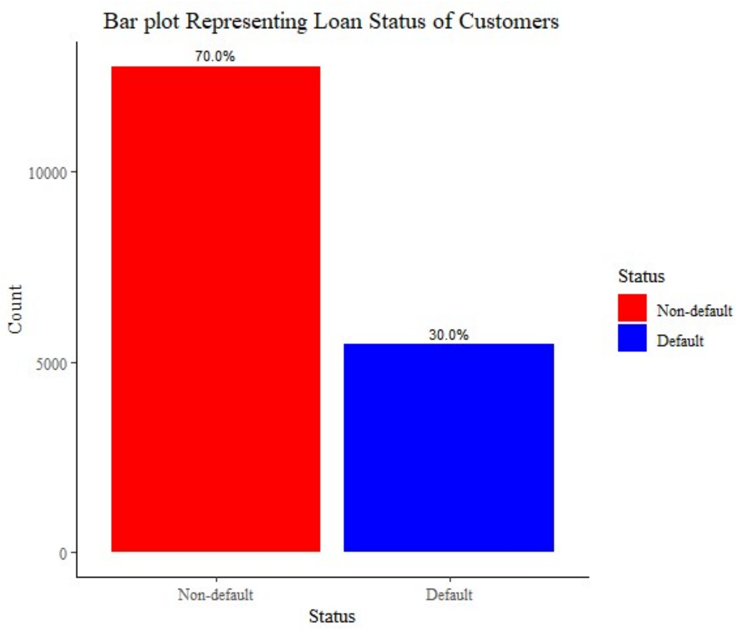

Figure 7: Bar Graph of Loan Characteristics

The total number of borrowers are 18196. Out of 18204, 70% did not default on their loans while 30% defaulted.

4.3.7 Credit History

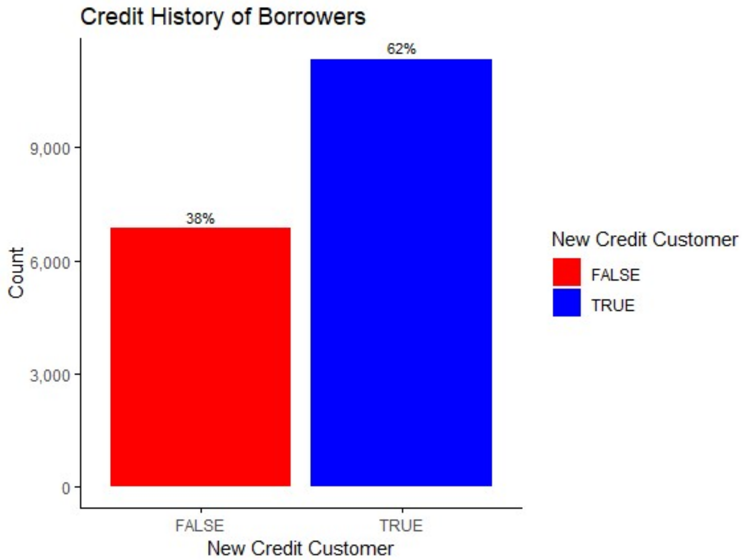

Figure 8: Bar Graph of Credit History

62% of the borrowers had no credit history while 38% of the borrowers had a credit history.

4.3.8 Interest

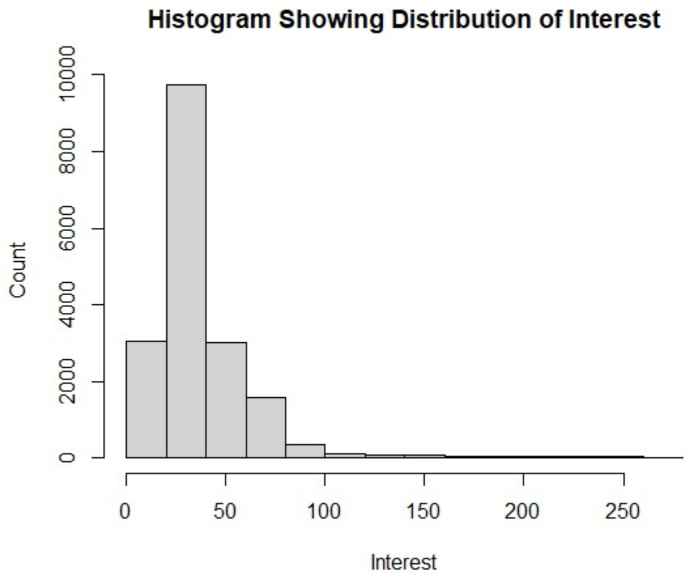

Figure 9: Histogram of Interests

The histogram is rightly skewed implying that a large number of loan applications had low interest rates.

4.3.9 Loan Amount

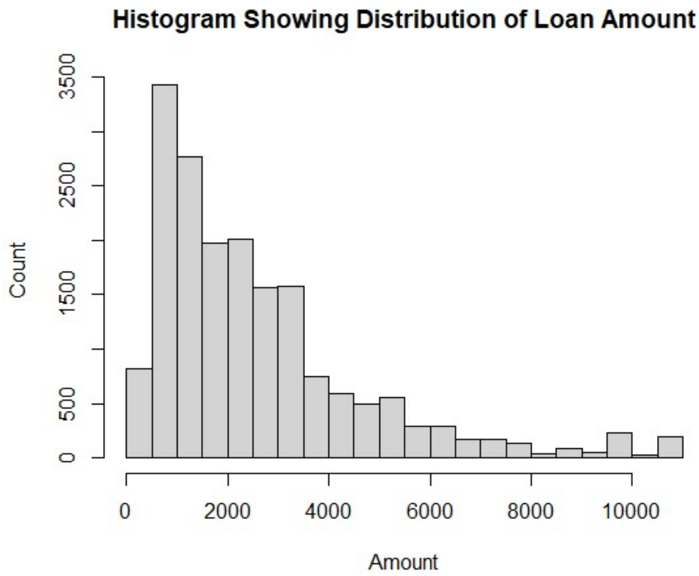

Figure 10: Histogram of Loan amount

This histogram indicates that most borrowers opted for smaller loan amounts. Most of the loans fall between the range of 500 euros to 3000 euros.

4.3.10 Use of Loan

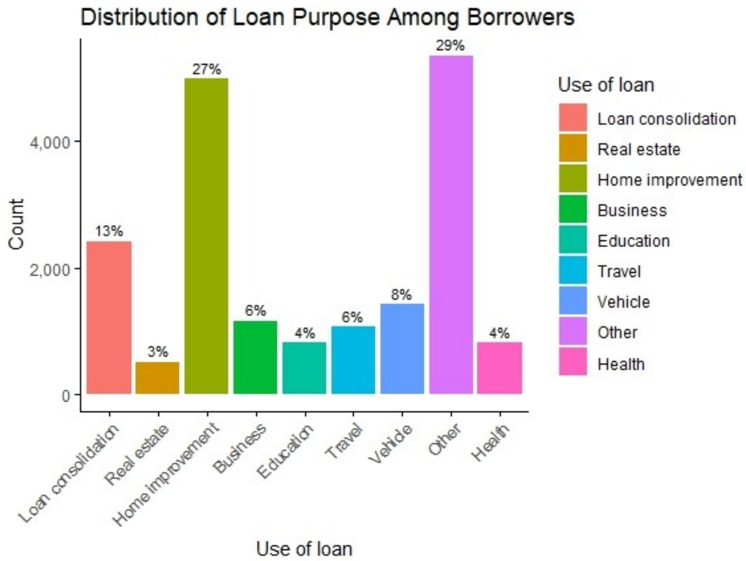

Figure 11: Bar Graph of Use of Loan

The majority of the borrowers used their loans for home improvement and other uses which were not directly specified while the least number of borrowers used their loans for real estate purposes.

4.3.11 Amount of Previous Loans

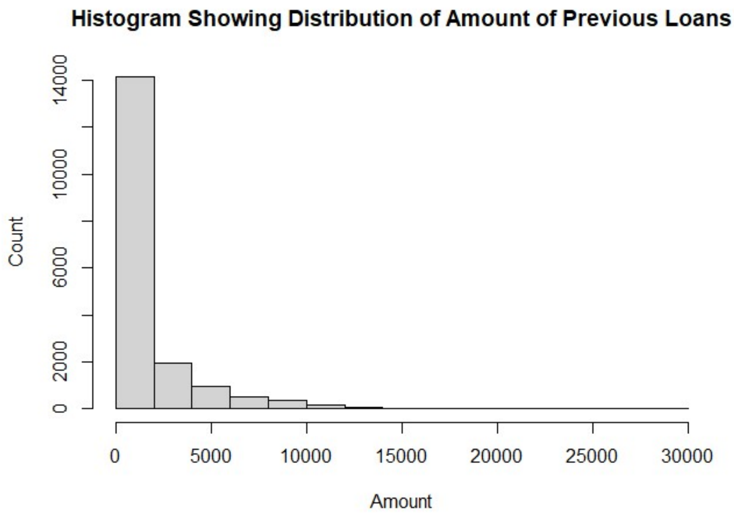

Figure 12: Histogran of Amount of Previous Loans

The plot of amount of previous loans is rightly skewed implying that the number of borrowers with less amount of previous loans are more than borrowers with larger previous loans.

4.4 Bivariate analysis

Bivariate analysis involves evaluating the relationship between two variables to understand how they are related to likelihood of defaulting on a bank loan.

	Non-default (N=12751)	Default (N=5453)	Overall (N=18204)
Loan_Data$GendCat			
Male	7129 (55.9%)	2760 (50.6%)	9889 (54.3%)
Female	5255 (41.2%)	1964 (36.0%)	7219 (39.7%)
Undefined	367 (2.9%)	729 (13.4%)	1096 (6.0%)
Loan_Data$agecat			
20 to 30	4101 (32.2%)	1160 (21.3%)	5261 (28.9%)
31 to 40	4007 (31.4%)	1779 (32.6%)	5786 (31.8%)
41 to 50	2683 (21.0%)	1389 (25.5%)	4072 (22.4%)
51 to 60	1435 (11.3%)	826 (15.1%)	2261 (12.4%)
61 to 70	525 (4.1%)	299 (5.5%)	824 (4.5%)
Loan_Data$CountryCat			
EE	8391 (65.8%)	767 (14.1%)	9158 (50.3%)
ES	2103 (16.5%)	2725 (50.0%)	4828 (26.5%)
FI	2254 (17.7%)	1961 (36.0%)	4215 (23.2%)
SK	3 (0.0%)	0 (0%)	3 (0.0%)
Loan_Data$UseofLoanCat			
Loan consolidation	1651 (12.9%)	738 (13.5%)	2389 (13.1%)
Real estate	374 (2.9%)	120 (2.2%)	494 (2.7%)
Home improvement	3507 (27.5%)	1367 (25.1%)	4874 (26.8%)
Business	849 (6.7%)	295 (5.4%)	1144 (6.3%)
Education	553 (4.3%)	245 (4.5%)	798 (4.4%)
Travel	725 (5.7%)	326 (6.0%)	1051 (5.8%)
Vehicle	1049 (8.2%)	363 (6.7%)	1412 (7.8%)
Other	3496 (27.4%)	1756 (32.2%)	5252 (28.9%)
Health	547 (4.3%)	243 (4.5%)	790 (4.3%)
Loan_Data$EduCat			
Primary Education	99 (0.8%)	112 (2.1%)	211 (1.2%)
Basic Education	1688 (13.2%)	664 (12.2%)	2352 (12.9%)
Vocational Education	2512 (19.7%)	1692 (31.0%)	4204 (23.1%)
Secondary Education	5115 (40.1%)	1569 (28.8%)	6684 (36.7%)
Higher Education	3337 (26.2%)	1416 (26.0%)	4753 (26.1%)
Loan_Data$MaritalStatusCat			
Married	3691 (28.9%)	1689 (31.0%)	5380 (29.6%)
Cohabitant	3339 (26.2%)	873 (16.0%)	4212 (23.1%)
Single	4417 (34.6%)	2142 (39.3%)	6559 (36.0%)
Divorced	1123 (8.8%)	665 (12.2%)	1788 (9.8%)
Widow	181 (1.4%)	84 (1.5%)	265 (1.5%)
Loan_Data$EmpStatusCat			
Partially employed	389 (3.1%)	237 (4.3%)	626 (3.4%)
Fully employed	10390 (81.5%)	4206 (77.1%)	14596 (80.2%)
Self-employed	453 (3.6%)	351 (6.4%)	804 (4.4%)
Entrepreneur	929 (7.3%)	193 (3.5%)	1122 (6.2%)

Table 1: Bivariate Analysis Table

	Non-default (N=12751)	Default (N=5453)	Overall (N=18204)
Retiree	590 (4.6%)	466 (8.5%)	1056 (5.8%)
Loan_Data$EmpDurCat			
TrialPeriod	4978 (39.0%)	2359 (43.3%)	7337 (40.3%)
UpTo1Year	2369 (18.6%)	916 (16.8%)	3285 (18.0%)
UpTo2Years	1815 (14.2%)	690 (12.7%)	2505 (13.8%)
UpTo3YearS	1531 (12.0%)	539 (9.9%)	2070 (11.4%)
UpTo4Years	994 (7.8%)	458 (8.4%)	1452 (8.0%)
UpTo5Years	1063 (8.3%)	490 (9.0%)	1553 (8.5%)
Missing	1 (0.0%)	1 (0.0%)	2 (0.0%)
Loan_Data$WorkExpCat			
LessThan2Years	567 (4.4%)	157 (2.9%)	724 (4.0%)
2To5Years	1781 (14.0%)	581 (10.7%)	2362 (13.0%)
5To10Years	2824 (22.1%)	1009 (18.5%)	3833 (21.1%)
10To15Years	2522 (19.8%)	1134 (20.8%)	3656 (20.1%)
15To25Years	2812 (22.1%)	1333 (24.4%)	4145 (22.8%)
MoreThan25Years	2245 (17.6%)	1238 (22.7%)	3483 (19.1%)
Missing	0 (0%)	1 (0.0%)	1 (0.0%)
Loan_Data$OccupCat			
Other	2741 (21.5%)	1375 (25.2%)	4116 (22.6%)
Mining	55 (0.4%)	19 (0.3%)	74 (0.4%)
Processing	1117 (8.8%)	211 (3.9%)	1328 (7.3%)
Energy	219 (1.7%)	97 (1.8%)	316 (1.7%)
Utilities	106 (0.8%)	80 (1.5%)	186 (1.0%)
Construction	1301 (10.2%)	443 (8.1%)	1744 (9.6%)
Retail and wholesale	1265 (9.9%)	462 (8.5%)	1727 (9.5%)
Transport and warehousing	913 (7.2%)	411 (7.5%)	1324 (7.3%)
Hospitality and catering	774 (6.1%)	508 (9.3%)	1282 (7.0%)
Info and telecom	706 (5.5%)	291 (5.3%)	997 (5.5%)
Finance and Insurance	415 (3.3%)	129 (2.4%)	544 (3.0%)
Real-estate	184 (1.4%)	102 (1.9%)	286 (1.6%)
Research	247 (1.9%)	55 (1.0%)	302 (1.7%)
Administrative	262 (2.1%)	218 (4.0%)	480 (2.6%)
Civil service & military	574 (4.5%)	214 (3.9%)	788 (4.3%)
Education	480 (3.8%)	136 (2.5%)	616 (3.4%)
Healthcare and social help	805 (6.3%)	511 (9.4%)	1316 (7.2%)
Art and entertainment	225 (1.8%)	79 (1.4%)	304 (1.7%)
Agriculture,forestry and fishing	361 (2.8%)	111 (2.0%)	472 (2.6%)
Loan_Data$HomeOwnCat			
Owner	4382 (34.4%)	1511 (27.7%)	5893 (32.4%)
Living with parents	2175 (17.1%)	1095 (20.1%)	3270 (18.0%)
Tenant,pre-furnished property	1834 (14.4%)	896 (16.4%)	2730 (15.0%)
Tenant,unfurnished property	1349 (10.6%)	882 (16.2%)	2231 (12.3%)

Table 2: Bivariate Analysis Table

	Non-default (N=12751)	Default (N=5453)	Overall (N=18204)
Council house	194 (1.5%)	139 (2.5%)	333 (1.8%)
Joint tenant	491 (3.9%)	174 (3.2%)	665 (3.7%)
Joint ownership	871 (6.8%)	234 (4.3%)	1105 (6.1%)
Mortgage	1260 (9.9%)	495 (9.1%)	1755 (9.6%)
Owner with encumbrance	195 (1.5%)	27 (0.5%)	222 (1.2%)

Table 3: Bivariate Analysis Table

The following were key observations observed during the analysis:

- The male customers defaulted more compared to the female customers.
- Customers in the age bracket 31 to 40 defaulted more compared to the other age groups.
- Majority of the loan defaulters were those who had attained vocational level of education.
- Most of the loan defaulters were single
- A majority of the loan defaulters are fully employed.
- Most of the loan defaulters had between 15 to 25 years of work experience.
- A majority of the loan defaulters were home owners.

4.5 Correlation Analysis

This plot shows variables that are highly correlated to each other.

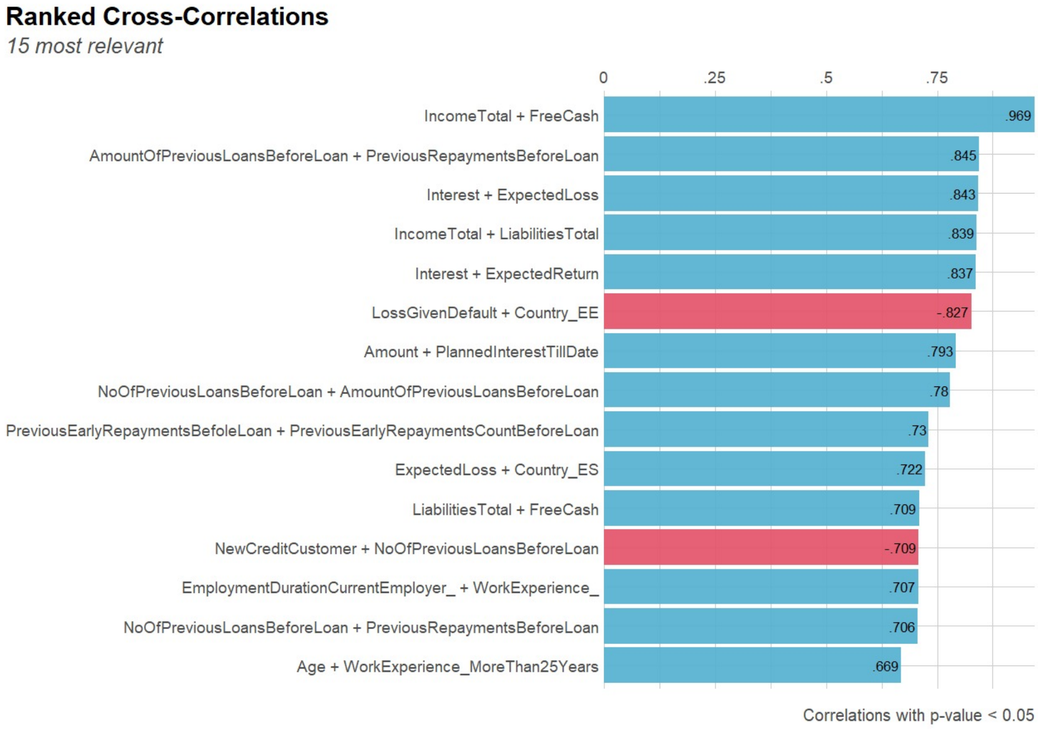

Figure 13: Correlation Analysis

A threshold of 0.05 was set to determine and remove highly correlated variables to reduce redundancy and improve the model efficiency. The following variables were removed after performing the correlation analysis; Free cash, Number of previous loans before loan, Previous repayments before loan, expected loss, expected return, liabilities total, country, Planned Interest till date, Previous early repayments count before loan, Work experience.

4.6 Logistic Regression Model

4.6.1 Data Transformation

The outcome variable was redefined to represent whether a loan was defaulted or not. Loans classified as late were considered defaults and assigned the value 1 while loans classified as repaid were considered non-defaults and assigned the value 0. Loans classified as current were foregone altogether since these represented loans that were still being serviced.

Loan number and loan date were also removed since loan number is a unique ID given to all loan applicants hence not useful for the study.

The New Credit Customer variable was transformed into a numerical variable where TRUE was represented by the value 1 and FALSE represented by 0.

Variables such as Gender,Education ,Marital status.Employment Status,Occupation area and Home ownership type were transformed into categorical variables to make it easier to understand the effects of different categories of a variable on loan default.

4.6.2 Splitting the data

The dataset was split in the ratio 2:1 where two thirds was used in training the model and the remaining one third was used in testing the model. This was done to ensure there was unbiased model performance.

4.6.3 Fitting the Logistic Regression Model

A logistic regression model was fit using the training set with all the predictor variables included. Backward stepwise logistic regression was then applied to the model to select the most significant variables. Wald's test was used to determine the significance of individual coefficients of the model.Likelihood ratio test was used to determine the significance of the model.

4.7 Interpretation of Model Coefficients

The logistic regression model provides insights into how different variables influence the likelihood of loan default. By interpreting the coefficients of the model, we can identify which factors significantly increase or decrease the risk of default. Since defaults were coded as 1 and non-defaults as zero the customers who did not default made up the reference group and the customers who defaulted constituted the group of interest. The following table represents the summary output of the model together with the coefficients, standard error, z values and p values.

```
Call:
glm(formula = Status ~ NCC + Age + GendCat + Amount + Interest +
    LoanDuration + MonthlyPayment + EduCat + MaritalStatusCat +
    EmpStatusCat + OccupCat + HomeOwnCat + DebtToIncome + PlannedPrincipalTillDate +
    LossGivenDefault + AmountOfPreviousLoansBeforeLoan + PreviousEarlyRepaymentsBefoleLoan,
    family = binomial, data = training_set)

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                 -7.655e+00  3.344e-01 -22.890  < 2e-16 ***
NCCTRUE                                      1.574e-01  6.974e-02   2.257 0.024023 *  
Age                                          2.415e-02  2.629e-03   9.187  < 2e-16 ***
GendCatFemale                                1.028e-01  5.560e-02   1.849 0.064409 .  
GendCatUndefined                             7.228e-01  1.125e-01   6.425 1.32e-10 ***
Amount                                       1.798e-04  1.616e-05  11.123  < 2e-16 ***
Interest                                     1.195e-02  1.337e-03   8.941  < 2e-16 ***
LoanDuration                                 7.818e-03  1.616e-03   4.839 1.31e-06 ***
MonthlyPayment                               1.432e-03  2.238e-04   6.400 1.55e-10 ***
EduCatBasic Education                       -1.037e-01  2.058e-01  -0.504 0.614435    
EduCatVocational Education                   6.365e-02  1.998e-01   0.319 0.750096    
EduCatSecondary Education                   -2.123e-01  1.997e-01  -1.063 0.287688    
EduCatHigher Education                      -3.294e-01  2.014e-01  -1.636 0.101889    
MaritalStatusCatCohabitant                  -1.037e-01  7.269e-02  -1.427 0.153644    
MaritalStatusCatSingle                       6.718e-04  6.532e-02   0.010 0.991794    
MaritalStatusCatDivorced                     6.226e-03  8.499e-02   0.073 0.941607    
MaritalStatusCatWidow                       -6.188e-01  2.048e-01  -3.022 0.002514 ** 
EmpStatusCatFully employed                  -3.167e-01  1.242e-01  -2.550 0.010781 *  
EmpStatusCatSelf-employed                   -4.634e-02  1.597e-01  -0.290 0.771627    
EmpStatusCatEntrepreneur                    -4.404e-01  1.652e-01  -2.667 0.007664 ** 
EmpStatusCatRetiree                         -1.161e-01  1.577e-01  -0.736 0.461712    
OccupCatMining                              -3.515e-01  3.686e-01  -0.954 0.340335    
OccupCatProcessing                          -2.513e-01  1.183e-01  -2.123 0.033719 *  
OccupCatEnergy                              -1.439e-01  1.961e-01  -0.734 0.462956    
OccupCatUtilities                            4.798e-01  2.237e-01   2.144 0.031993 *  
```

Figure 14: Model Summary

```
OccupCatConstruction                       -4.432e-02  9.826e-02  -0.451 0.651952
OccupCatRetail and wholesale                9.110e-02  9.383e-02   0.971 0.331574
OccupCatTransport and warehousing           4.377e-02  1.015e-01   0.431 0.666320
OccupCatHospitality and catering            2.079e-01  1.011e-01   2.056 0.039775 *
OccupCatInfo and telecom                   -1.585e-01  1.153e-01  -1.375 0.169184
OccupCatFinance and Insurance              -4.259e-01  1.549e-01  -2.749 0.005976 **
OccupCatReal-estate                         2.272e-01  1.805e-01   1.259 0.208079
OccupCatResearch                           -2.941e-01  1.989e-01  -1.479 0.139188
OccupCatAdministrative                     -3.354e-02  1.423e-01  -0.236 0.813618
OccupCatCivil service & military           -2.929e-01  1.243e-01  -2.356 0.018486 *
OccupCatEducation                          -3.435e-01  1.521e-01  -2.258 0.023917 *
OccupCatHealthcare and social help          8.227e-02  9.792e-02   0.840 0.400819
OccupCatArt and entertainment               1.012e-01  1.923e-01   0.526 0.598693
OccupCatAgriculture,forestry and fishing    1.201e-01  1.597e-01   0.752 0.451862
HomeOwnCatLiving with parents               2.906e-01  7.825e-02   3.714 0.000204 ***
HomeOwnCatTenant,pre-furnished property     2.224e-01  7.851e-02   2.833 0.004612 **
HomeOwnCatTenant,unfurnished property       3.209e-01  7.870e-02   4.077 4.56e-05 ***
HomeOwnCatCouncil house                     3.542e-01  1.669e-01   2.123 0.033762 *
HomeOwnCatJoint tenant                      2.229e-01  1.378e-01   1.618 0.105765
HomeOwnCatJoint ownership                  -2.775e-02  1.133e-01  -0.245 0.806494
HomeOwnCatMortgage                         -1.167e-01  8.814e-02  -1.324 0.185611
HomeOwnCatOwner with encumbrance           -7.981e-01  2.930e-01  -2.724 0.006456 **
DebtToIncome                                2.305e-03  1.439e-03   1.602 0.109117
PlannedPrincipalTillDate                   -5.293e-04  2.884e-05 -18.349  < 2e-16 ***
LossGivenDefault                            6.716e+00  2.262e-01  29.689  < 2e-16 ***
AmountOfPreviousLoansBeforeLoan             4.596e-05  1.426e-05   3.223 0.001268 **
PreviousEarlyRepaymentsBefoleLoan          -3.174e-05  1.401e-05  -2.265 0.023527 *
---
```

Figure 15: Model Summary

The model coefficients were exponentiated to form the column of odds ratio to ease interpretation.

Variables	Odds Ratio
Intercept	0.0004
Age	1.0244
Amount	1.0001
Interest	1.012
Loan Duration	1.0078
Marital Status-Widow	0.0539
Gender-Female	1.108
Employment Status-fully employed	0.7286
Employment Status-entrepreneur	0.6438
Occupation Area-finance and insurance	0.6532
Occupation Area-processing	0.7778
Occupation Area-utilities	1.616
Occupation Area-hospitality and catering	1.2311
Occupation Area-civil service and military	0.7461
Occupation Area-education	0.7093
Home ownership type-living with parents	1.3373
Home ownership type-tenant pre furnished	1.249
Home ownership type-tenant unfurnished	1.378
Home ownership type-council house	1.425
Home ownership type-owner with encumbrance	0.4501
Planned Principal till date	0.9995
Loss Given Default	0.0825
Amount of previous loans before loan	1.00005
Previous early repayments before loan	0.9

Table 4: Odds Ratio Table

Intercept

The intercept represents the estimated log odds of default when the independent variables are zero. For our model, the intercept has an odds ratio 0f 0.0004 implying that when the predictors are set to zero the odds of defaulting are minimal.

Age

The age variable has an odds ratio of 1.0244.For each one-unit increase in age, a borrower is 2.44% more likely to default

Amount

The amount variable has an odds ratio of 1.0001. For every one-unit increase in the loan amount, a borrower is 0.01% more likely to default.

Interest

The interest variable has an odds ratio of 1.012. For every one-unit increase in the interest rate, a borrower is 0.012% more likely to default.

Loan Duration

The loan duration variable has an odds ratio of 1.0078. For every one-unit increase in loan duration, a borrower is 0.78% more likely to default.

Marital Status.

The marital status variable has five categories: “Married”,”Cohabitant”, ”Single”, ”Divorced” and “Widow”. The “Married” category was used as the reference group. The odds ratio for the widow category was 0.0539 implying that individuals who are widowed were 94.61% less likely to default compared to those individuals who were married.

Gender

The “male” category was used as the reference group. The female category had an odds ratio of 1.108 implying that females were 10.8% more likely to default on their loans compared to males.

Employment Status

The employment status variable has five categories: “Partially employed”, ”Fully employed”, ”Self-employed”, ”Retired”, “Entrepreneur” . The “Partially employed” category was used as the reference group. The odds ratio for the fully employed category was 0.7286 implying that individuals who are fully employed were 27.14% less likely to default on their loans than those who were partially employed. The odds ratio for the entrepreneur category was 0.6438 implying that entrepreneurs were 35.62% less likely to default on their loans compared to those who were partially employed.

Occupation Area

The occupation area variable has 19 categories. The “Other” category was used as the reference group. The odds ratio for the category finance and insurance was 0.6532 implying that individuals in finance and insurance were 34.63% less likely to default than those whose professions were not specified. The odds ratio for the category processing was 0.7778 implying that individuals in processing were 22.22% less likely to default than those whose professions were not specified. The odds ratio for the category utilities was 1.616 implying that individuals in utilities were 61.6% more likely to default than those whose professions were not specified. The odds ratio for the category hospitality and catering was 1.2311 implying that individuals in hospitality and catering were 23.11% more likely to default than those whose professions were not specified. The odds ratio for the category Civil service and military was 0.7461 implying that individuals in civil service and military were 25.39% less likely to default than those whose professions were not specified. The odds ratio for the category education was 0.7093 implying that individuals in finance and insurance were 29.07% less likely to default than those whose professions were not specified.

Home Ownership Type

The home ownership category has 9 categories. The "Owner" category was used as the reference group. The odds ratio for the category living with parents was 1.3373 implying that individuals that were living with parents were 33.73% more likely to default than those who owned their properties. The odds ratio for the category tenant ,pre furnished was 1.249 implying that individuals that tenants,pre furnished were 24.9% more likely to default than those who owned their properties. The odds ratio for the category tenant, unfurnished property was 1.378 implying that individuals that were tenant, unfurnished were 37.8% more likely to default than those who owned their properties. The odds ratio for the category Council house was 1.425 implying that individuals that were living in council houses were 42.5% more likely to default than those who owned their properties. The odds ratio for the category owner with encumbrance was 0.4501 implying that individuals that were owners with encumbrance were 54.99% less likely to default than those who owned their properties

Planned Principal Till Date

The planned principal till date variable has an odds ratio of 0.9995. For every one-unit increase in the planned principal till date, a borrower is 0.05% less likely to default.

Loss Given Default

The loss given default variable has an odds ratio of 0.0825. For every one-unit increase in the loss given default amount, a borrower is 91.75% less likely to default.

Amount of Previous loans before loans

The amount of previous loans before loans variable has an odds ratio of 1.00005 .For every one-unit increase in the amount of previous loans before loan, a borrower is 0.005% more likely to default.

Previous Early Repayments before loan

The amount of previous early repayments before loan variable has an odds ratio of 0.9. For every one-unit increase in the amount of previous early repayments before loan, a borrower is 10% less likely to default.

4.8 Model Evaluation

4.8.1 Significance of Predictors(Wald's Test)

```
Wald test

Model 1: Status ~ NCC + Age + GendCat + Amount + Interest + LoanDuration +
    MonthlyPayment + EduCat + MaritalStatusCat + EmpStatusCat +
    OccupCat + HomeOwnCat + DebtToIncome + PlannedPrincipalTillDate +
    LossGivenDefault + AmountOfPreviousLoansBeforeLoan + PreviousEarlyRepaymentsBefoleLoan
Model 2: Status ~ 1
  Res.Df  Df      F    Pr(>F)
1  11620
2  11671 -51 43.449 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 16: Wald's Test analysis

Model 1 represents the full model containing all predictor variables while model 2 represents the reduced model that includes only the intercept. The Wald's test results suggests that model 1 provides a better fit to the data than model 2. The predictors included in model 1 are essential in predicting loan default probabilities. Given the extremely low p-value

$$2.2e^{-16} < 0.001$$

, we reject the null hypothesis that the additional predictors in model 1 are not necessary.

4.8.2 Significance of the Model(Likelihood Ratio Test)

```
Likelihood ratio test

Model 1: Status ~ NCC + Age + GendCat + Amount + Interest + LoanDuration +
    MonthlyPayment + EduCat + MaritalStatusCat + EmpStatusCat +
    OccupCat + HomeOwnCat + DebtToIncome + PlannedPrincipalTillDate +
    LossGivenDefault + AmountOfPreviousLoansBeforeLoan + PreviousEarlyRepaymentsBefoleLoan
Model 2: Status ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  52 -5460.7
2   1 -7063.0 -51 3204.7  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 17: Likelihood Ratio test

Model 1 represents the full model containing all predictor variables while model 2 represents the reduced model that includes only the intercept. With a p-value $2.2e^{-16} < 0.001$, the null hypothesis that the reduced model is a better fit is rejected at 0.001 level of significance. The likelihood ratio test results indicates that model 1 provides significantly better fit to the data than model 2.

4.8.3 Confusion Matrix

The model's performance was evaluated using a confusion matrix. Out of 5836 predicted values, there were 2062 True negative(Loans that were predicted to not default and actually did not default), 1929 False positives(Loans that were predicted to default but did not default), 126 False negatives(Loans that were predicted to not default but defaulted), 1719 True positives(Loans that were predicted to default and defaulted)

	Predicted-No default	Predicted-Default
Actual-No default	2062	1929
Actual-default	126	1719

Table 5: Confusion Matrix

Accuracy

The model's accuracy was 65% which means that 65% of the total predictions(both positive and negative) were correctly classified. This represents the overall effectiveness of the model in differentiating between defaulters and non-defaulters.

Sensitivity

The model's sensitivity was 93% which is the percentage of defaulters that were classified correctly. The high sensitivity is important in predicting loan default as it ensures most defaulters are accurately detected thus reducing the risk of granting loans to high-risk borrowers.

Specificity

The model's specificity was 52% which is the percentage of non-defaulters that were classified correctly. Whereas the percentage is lower than desired, it still shows the model's ability to correctly identify borrowers who are non-defaulters.

5 Chapter 5:Summary,Conclusions and Recommendations

5.1 Introduction

This chapter gives the summary, conclusions and recommendations drawn from the findings.

5.2 Summary

The main objective of this study was to use logistic regression to model the relationship between various independent variables and loan default outcomes. Through logistic regression analysis, the specific objectives were met by identification of key demographic and socio- economic factors and loan specific characteristics associated with bank loan default.

5.3 Conclusions

Based on the analysis of the dataset, several significant insights into the impact of demographic, socio-economic and loan-specific characteristics in predicting loan default were identified. Results from the bivariate analysis show that certain categories including male borrowers, individuals aged 31 to 40, borrowers who had attained vocational level of education, borrowers who were single, fully employed borrowers and homeowners were more likely to default on their loans. In addition from the logistic regression model loan specific characteristics such as amount, interest, loan duration, monthly payment, planned principal till date, loss given default, previous early repayments before loan and the amount of previous loans before loan were significant in predicting probability of default. The model evaluation showed a low specificity of 52%, a high sensitivity of 93% and an average accuracy of 65%. This revealed that although the model performs well in identifying borrowers who default, it may have limitations in correctly pointing out non -defaulters. In conclusion, this study highlights the importance of taking into account both demographic and loan specific characteristics in predicting bank loan defaults. By maximizing on these findings, financial institutions can better assess the risk associated with lending thus enhancing decision making procedures and lowering potential losses.

5.4 Recommendations for Further Research

- Our model was not fully accurate therefore, we recommend use of advanced machine learning techniques to be explored to enhance model accuracy and interpretability.
- This study explored demographic, socio-economic and loan specific characteristics, further research could incorporate variables such as co-signer or guarantor , health and lifestyle factors and political and economic indicators to enhance the accuracy of credit risk assessment.
- This study used quantitative data, further research could benefit from including qualitative research methods such as surveys or interviews with loan borrowers and credit officers to gain deeper understanding of the underlying causes of loan default.

6 References

- Abid, L., Ouertani, M. N., amp; Zouari-Ghorbel, S. (2014). Macroeconomic and Bank-specific Determinants of Household's Non-performing Loans in Tunisia: A Dynamic Panel Data. Procedia Economics and Finance, 13, 58–68. https://doi.org/10.1016/s 5671(14)00430-4

- Abras, Ana, and Bruno de Paula Rocha. "Bank Credit Shocks and Employment Growth: An Empirical Framework for the Case of Brazil." Journal of Developing Areas, vol. 54, no. 1, 2020, pp. 1–18,

- Addae-Korankye, A. (2014). Causes and Control of Loan Default/ Delinquency in Microfinance Institutions in Ghana. American International Journal of Contemporary Research, 4(12).

- Aslam, U., Tariq Aziz, H. I., Sohail, A., Batcha, N. K. (2019). An Empirical Study on Loan Default Prediction Models. Journal of Computational and Theoretical Nanoscience, 16(8), 3483–3488.

  https://doi.org/10.1166/jctn.2019.8312

- Banton, C. (2023, March 28). Interest rates: Different types and what they mean to borrowers.Investopedia.

  https://www.investopedia.com/terms/i/interestrate.asp

- Beaton, Kimberly , et al. "International Monetary Fund." Www.imf.org, Nov. 2016,

- De Bock, Reinout, and Alexander Demyanets. "IMF Working Papers Volume 2012 Issue 071: Bank Asset Quality in Emerging Markets: Determinants and Spillovers (2012)." IMF ELibrary, 12 Mar. 2012,

  www.elibrary.imf.org/view/journals/001/2012/071/001.2012.issue-071-en.xml.

- Eisenhardt, K.M. (1989). Agency theory: an assessment and review. Academy of Management Review, 14, 57-74.

- Ekanayake, S. (2004). Agency theory, national culture and management control systems. The Journal of American Academy of Business Cambridge, 4(1/2), 49-54.

- EQUITY BANK (KENYA) LIMITED. (n.d.). Retrieved March 27, 2024, from https://equitygroupholdings.com/wp-content/uploads/2022/08/Equity-Group-Holdings-PLC- Financial-Statements-for-the-Period-ended-30th-June-2022.pdf
- Eric Maina, G. (2014). Assessing Institutional Factors Contributing To Loan Defaulting In Microfinance Institutions in Kenya. IOSR Journal of Humanities and Social Science, 19(5), 105–123.

  https://doi.org/10.9790/0837-1952105123
- European Central Bank. (2018, March 25). Papers by Claudia Schwarz. European Central Bank.

  https://www.ecb.europa.eu/pub/research/authors/profiles/claudia-schwarz.en.html
- Jabbouri, Imad, and Maryem Naili. "Determinants of Nonperforming Loans in Emerging Markets: Evidence from the MENA Region." Review of Pacific Basin Financial Markets and Policies (RPBFMP), vol. 22,no. 04, 2019, pp. 1–33,
- James, K. (2019). DETERMINANTS OF LOAN REPAYMENT DEFAULTS IN MICROFINANCE BANKS IN KENYA. International Journal of Social Sciences Management and Entrepreneurship (IJSSME), 3(1).

  https://sagepublishers.com/index.php/ijssme/article/view/36
- January 24 2024, W. (2024, January 24). Loan defaults up 26 percent to Sh634bn. Business Daily.

  https://www.businessdailyafrica.com/bd/economy/loan-defaults-up-26-percent-to-sh634bn--4502180
- Jensen, M. C., Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of financial economics, 3(4), 305-360.
- Kenya lenders plan to step up asset seizures over bad loans. (2024, February 24). The East African.

  https://www.theeastafrican.co.ke/tea/business/kenya-lenders-

  plan-to-step-up-asset-seizures-over- bad-loans-4535776

- Klein, Nir. "Non-Performing Loans in CESEE: Determinants and Impact on Macroeconomic Performance."
- Lin, T. T., Lee, C.-C., amp; Chen, C.-H. (2011). Impacts of the borrower's attributes, loan contract contents, and collateral characteristics on mortgage loan default. The Service Industries Journal, 31(9), 1385–1404.

  https://doi.org/10.1080/02642060903437535
- Louzis, D. P., Vouldis, A. T., amp; Metaxas, V. L. (2012). Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios. Journal of Banking amp; Finance, 36(4), 1012–1027.
- Mabeya, J. (2019). Determinants Of Loan Repayment Default Among Small And Medium Enterprises In Kisumu Central Sub County, Kenya. Erepository.uonbi.ac.ke.

  http://erepository.uonbi.ac.ke/handle/11295/107796
- Mishkin, F. S. (2007). The economics of money, banking, and financial markets. Pearson Education India.
- Nawai, N., amp; Shariff, M. N. M. (2012). Factors Affecting Repayment Performance in Microfinance Programs in Malaysia. Procedia - Social and Behavioral Sciences, 62, 806–811.

  https://doi.org/10.1016/j.sbspro.2012.09.136
- Netzer, O., Lemaire, A., amp; Herzenstein, M. (2019). When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. Journal of Marketing Research, 56(6), 960–980.

  https://doi.org/10.1177/0022243719852959
- Nigmonov, A., Shams, S., amp; Alam, K. (2022). Macroeconomic determinants of loan defaults: Evidence from the U.S. peer-to-peer lending market. Research in International Business and Finance, 59, 101516.

  https://doi.org/10.1016/j.ribaf.2021.101516

- Ranisavljević, D., Hadžić, M. (2016). Realistic Evaluation of the Ratio: Loan-To-value – The Key to Minimising the Credit Risk. Economic Themes, 54(3), 449–468. https://doi.org/10.1515/ethemes- 2016-0022
- Shefrin, H. M. (2002). Behavioral corporate finance. Oxford University Press
- Shepard, D. (2022, August 31). What's the Purpose of a Loan? Reasons to Get a Personal Loan. LendingTree. https://www.lendingtree.com/personal/8-good-reasons-to-get-a-personal-loan/
- Vandell, K. D., amp; Thibodeau, T. (1985). Estimation of Mortgage Defaults Using Disaggregate Loan History Data. Real Estate Economics, 13(3), 292–316. https://doi.org/10.1111/1540- 6229.00356
- Yuksel, Serhat, and Sinemis Zengin. "Influencing Factors of Net Interest Margin in Turkish Banking Sector." International Journal of Economics and Financial Issues, vol. 7, no. 1, 2017, pp. 178–191
- Zhang, L., Wang, J., Liu, Z. (2023). What should lenders be more concerned about? Developing a profit-driven loan default prediction model. Expert Systems with Applications, 213, 118938. https://doi.org/10.1016/j.eswa.2022.118938