UNIVERSITY OF CAPE COAST

MODELLING LOAN-DEFAULTING TENDENCIES AMONG CUSTOMERS OF A LOCAL GHANAIAN BANK

MICHAEL KOFI ASARE

2023

UNIVERSITY OF CAPE COAST

MODELLING LOAN-DEFAULTING TENDENCIES AMONG CUSTOMERS OF A LOCAL GHANAIAN BANK

BY

MICHAEL KOFI ASARE

Thesis submitted to the Department of Statistics of the College of Agriculture and Natural Sciences, School of Physical Sciences, University of Cape Coast, in partial fulfilment of the requirements for the award of Doctor of Philosophy degree in Statistics

MAY 2023

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree at this university or elsewhere.

Candidate's Signature ……Kofioforiasare…… Date …21/07/2023…

Name: Michael Kofi Asare

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis was supervised in accordance with guidelines on supervision of thesis laid down by the University of Cape Coast.

Supervisor's Signature……………………… Date…………………

Name: Prof. Nathaniel Howard

Co-supervisor's Signature……………………… Date…………………

Name: Dr. Francis Eyiah-Bediako

ABSTRACT

In Ghana, some local banks lack a comprehensive credit risk management framework that includes the use of credit scoring; hence, the purpose of this study is to develop a credit scoring tool derived from a binary logistic regression model to reduce credit risk exposure for banks in general and local banks in particular. A review of the literature on credit scoring models or classifiers revealed that the specificity and sensitivity of the developed models are not explored further to reveal insights into the optimal cut-off point of the model. This study seeks to fill this gap by further exploring the specificity and sensitivity of the developed model and offers explanations and insights about the variations of the optimal cut-off point. The study makes a case for using the optimal cut-off point as a practical decision point for financial institutions. Secondary data on borrowers were obtained from a local bank and examined to identify its retail customers' demographic and behavioural characteristics based on the minimum Know Your Customer (KYC) required by the central bank. The binary logistic regression model developed from the data had an overall classification accuracy of 93% at a cut-off point of 0.5; however, the sensitivity measure was barely 23%. Typically, resampling techniques are employed to deal with the imbalance, however, in this study a plot of sensitivity and specificity against the probability of default is used to derive an optimal cut-off point. The performance of the logistic regression model at the derived optimal cut-off point was found to be similar to other binary models' performance that used resampling techniques.

KEY WORDS

Binary logistic regression

Credit risk management

Credit scoring

Optimal cut-off point

Sensitivity

Specificity

ACKNOWLEDGEMENTS

I would like to thank Professor Nathaniel Howard, whose advice and mentoring have been crucial throughout my research path. I would also like to thank the bank manager who provided the essential data for the study.

I am grateful to all my instructors and lecturers for their continuous encouragement and support. Their views and knowledge have influenced the course of my study.

Finally, I would want to convey my heartfelt gratitude to the individuals that helped my thesis come to fruition in various ways. Their help was critical in guaranteeing the quality and completion of this research.

DEDICATION

To my parents, Samuel, and Seraphina Asare, I attribute the inspiration behind this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AQR	Asset Quality Review
AUC	Area Under the Curve
BoG	Bank of Ghana
C&R	Classification and Regression Tree
CBG	Consolidated Bank Ghana LTD
CHAID	Chi-squared Automatic Interaction Detection
CRR	Cash Reserve Ratio
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSVM	Linear Support Vector Machine
MLRM	Multiple Logistic Regression Model
NPL	Non-performing Loans
QUEST	Quick, Unbiased, Efficient Statistical Trees
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

CHAPTER ONE

INTRODUCTION

This chapter introduces the study by first discussing the background and context, the causes of the problem, existing solutions, and the definition of key terms. Subsequently, the chapter provides the statement of the problem, the research objectives, the significance of the study and its limitations, and finally, the structure of the study.

Background to the Study

Monday, August 14, 2017, will go down in infamy in the minds of customers who held deposits with two of the nation's indigenous banks, UT Bank and Capital Bank. On that day, the Bank of Ghana's (BoG) initiative to clean up the financial system was launched in front of the public.

It all began in 2016 with a BoG Asset Quality Review (AQR) program, which uncovered various challenges plaguing Ghana's banking sector. These challenges included lack of capital, increased non-performing loans (NPLs) resulting from weak liquidity and credit risk management, and the absence of strong corporate governance frameworks. Additionally, the AQR exercise found that most of Ghana's indigenous banks were sensitive to the above mentioned challenges.

BoG liquidated two banks (Capital Bank and UT Bank) in August 2017 to avert a total financial sector collapse. A year after, seven more banks were declared insolvent and merged to constitute the Consolidated Bank Ghana (CBG), following the failure of several attempts to resuscitate these failing banks (Bank of Ghana, 2018). As a result of the clean-up, 9 universal banks, 347 micro-finance companies, 39 microcredit companies, 15 savings and loans companies, 8 financial house firms, and 2 non-banking institutions were closed (Bank of Ghana, 2019).

This study is motivated by the events described above in Ghana. Interestingly, all the closed banks are locally controlled commercial banks; thus, this study focuses on Ghana's locally controlled commercial banks.

Causes of the Problem

According to the BoG, reasons for the closure include insolvency resulting from capital inadequacy, questionable transactions, weak corporate governance frameworks and credit risk resulting in high levels of non-performing loans. The BoG identified credit risk as the chief contributor to the collapse of the locally controlled banks.

Research confirms the assertions of the BoG and has shown that the primary cause of serious banking challenges continues to be directly related to over-lenient or lax credit standards and poor lending practices by commercial banks (Bhanot, 2021; Gao & Zhu, 2020; Gupta & Bansal, 2020; Khan & Iqbal, 2019; Okafor & Onyekwelu, 2019). Lax credit standards and poor lending practices entail insider lending to high-risk borrowers. Other factors beyond the local banks' control are macroeconomic instability, lack of liquidity support and prudential regulations. The above mentioned factors were identified by the BOG as the bane of Ghana's locally controlled banks. These factors increase credit risk markedly and lead to NPLs occurrence. NPLs are bank loans that have been delinquent for more than a specified number of days, usually more than ninety or are deemed unlikely to be paid according to predefined criteria (Basel Committee on Banking Supervision, 2016). Credit or loans in this study refers to loans, overdrafts, mortgages, and trade financing.

The economy of a nation, especially developing nations, depends on all banks, whether they are locally owned or controlled by foreign entities. Creating capital for development, assisting the central bank in maintaining the value of the national currency, and, finally, adding value to or improving the lives of the nation's citizens are the three main roles or functions of banks. Banks create capital by using equity from their shareholders and deposits from the public. In more depth, the process of generating capital entails playing a part in the financial intermediation process, where banks take money from the economy's surplus side (in the form of current, call, savings accounts, and investment accounts) and lend it to the deficit side of the economy (in the form of loans, overdrafts, mortgages etc.). Not only do banks create credit to properly implement the financial intermediation process, but the banks also provide other essential services such as custodianship and acting as agents for their customers. The primary source of revenue for banks among all the services they offer is the generation of credit.

However, in performing the financial intermediation process, banks face many risks. The risks can broadly be categorised into three groups, credit risk, operational risk, and market risk (Allen & Saunders, 2012; Bellini, 2011; Matthews & Thompson, 2020; Saunders & Cornett, 2008). This study focuses on credit risk and proposes a statistical technique to mitigate it.

Credit Risk

With respect to direct losses, credit risk is the most noticeable and substantial risk in banking. In the worst-case scenario, a small number of significant clients going bad might cause a bank to suffer enormous losses and go bankrupt. Credit risk is the probability that loans will not be repaid or investments will lose value or default, causing a loss to the bank (Bessis, 2002). Credit risk involves the possibility that payments would be late, which could result in issues for the bank and the risk that borrowers could fail on their obligations.

As previously highlighted, locally controlled banks were found to engage in two key poor credit practices: insider lending and lending to high-risk borrowers. These practices, as mentioned above, in conjunction with the external factors, macroeconomic instability, lack of liquidity support, and prudential regulations, all come together and cause credit risk. We explain these factors briefly.

Insider lending

Insider lending was the most prevalent amongst the collapsed banks and was the single most significant contributor to the failure of the local banks. In at least half of the bank failures, insider credit facilities were responsible for a sizable share of the bad debts. For example, the BoG, in 2018, reported that one of the defunct banks, UNIBANK, had its shareholders, affiliate individuals, and organizations absorb a staggering GH¢5.3 billion, constituting 75% of the bank's total assets, resulting in a dramatic increase in their NPL portfolio. Insider lending threatened the locally controlled banks' soundness in that the credit facilities granted to the insiders were used to fund speculative ventures such as real estate development. Additionally, the credit facilities were used for non-returning projects (such as hotels and shopping centres), extending the maturities of the bank's resources and obligations. By engaging in insider lending, the banks surpassed their maximum exposure to huge credit facilities. The prevalence of insider lending in defunct banks implied that moral hazard issues were particularly prominent in those banks. We discuss some of the actions of the defunct banks that constitute a moral hazard.

Initially, individuals with no background in banking and finance, such as politicians, clergymen, chiefs, and queen mothers, were board members and directors of several local banks. Many deposits from the public sector were attracted via political connections. Many failed banks largely depended on large wholesale deposits from a select group of powerful people. Due to political pressure, it is unlikely that the influential individuals that made these deposits made a purely business judgment regarding their investments' safety. Furthermore, the availability of those deposits alleviated the need to compete for private sector deposits or financing. As a result, these banks were not under any depositor pressure to establish a trustworthy reputation.

Political ties have occasionally been used as leverage when applying for bank licenses. They were employed in some cases to convince the BoG not to punish banks for breaking banking regulations. All of the worries, as mentioned earlier, led to a relaxation of restrictions on dangerous banking activities. Due to loose management restrictions, banks came under pressure from their political connections to make loans to elected officials, board members, and related people and businesses in exchange for assistance obtaining deposits, licenses, and other things.

Furthermore, all failed banks had insufficient capital and could not reach the revised BoG capitalization threshold of GH¢400 million in 2017.

GH¢120 million served as the prior cap (Bank of Ghana, 2017). The prospective risks and advantages were drastically different under the previous capitalization cap because bank owners could lose only a small fraction of their capital if their banks failed. The bank's management was able to invest bank deposits in high-risk businesses since they knew they would earn greatly if the projects were a success and would suffer minimal losses if they failed. The weak local banks that the BoG purchased in 2018 had poor capital adequacy ratios and significant NPLs due to insider lending (Bank of Ghana, 2017).

The last factor that contributed to insider lending was an abnormal degree of ownership concentration. In every bank that failed, a single person or family owned the majority of the stock. These banks' management lacked appropriate independence from owner control in operational decisions (Elijah, 2019). Higher insider loan limitations may have been expected with a more diverse ownership structure and autonomous management.

Lending to high-risk borrowers

The second significant factor contributing to the local banks' collapse was lending to borrowers in high-risk parts of the credit market at high-interest rates. Lending to high-risk borrowers entails moral hazards on the side of both the banks and borrowers. It is triggered in part by the high expense of raising funding. Because depositors see local banks as less secure than established foreign banks, local banks are forced to give depositors higher interest rates. They also struggle to draw current accounts that do not pay interest because of their limited capacity to offer current account holders advantages over well-established foreign banks. Actual interest rates on borrowing from other banks and financial institutions by a few local banks were frequently higher than 20%. Due to the high cost of financing, local banks are forced to charge high lending rates, which negatively affect the quality of their loan portfolios and allow for a high return on assets. Local banks often face an unfavorable selection of borrowers since they may not meet the stringent credit worthiness criteria imposed on them. As a result, many borrowers who are turned down by foreign banks or would have been denied a loan tend to approach local banks. This leads to a higher proportion of borrowers with a higher risk of defaulting seeking loans from local banks. Local banks find it difficult to compete with foreign banks for the business of "prime" customers or the most creditworthy borrowers because they must charge higher interest rates on their loans to cover higher funding costs. As a result, credit markets are fragmented. Due to the lack of alternative credit sources, many local banks operate in the riskiest industry, offering services to customers willing to pay high-interest rates. Other banks and non-bank financial institutions (NBFIs) posed a high risk of default due to their lack of liquidity and willingness to accept interest rates above the market (Bank of Ghana, 2019).

The volume of interbank lending among local banks caused bank problems to cascade. There were both high-quality (i.e., creditworthy) borrowers and low-quality risks in the credit market segments that local banks managed. However, due to severe informational inadequacies, servicing consumers in this market requires comprehensive loan appraisal and monitoring systems: Because many borrowers lack experience running profitable enterprises, the quality of the borrowers' financial accounts is typically insufficient. Many failed banks struggled to identify good and bad risks because they lacked experience screening and monitoring their borrowers. Internal controls and credit procedures, including documentation for loans and loan security, were also frequently lacking. These institutions' managers and directors frequently lacked the necessary abilities and information. Since larger banks can offer the most talented bank officials better career opportunities, local banks frequently struggle to find competent employees.

Macroeconomic instability

The quality of local banks' loans is significantly impacted by macroeconomic uncertainty in two ways. Due to its unpredictable character and the significant degree of variation in the rates of increase of the prices of the particular goods and services that make up the total price index, high inflation has two main effects. The first is that it makes corporate earnings more erratic. Businesses are more likely to experience losses and to benefit unexpectedly, both of which are increasing probabilities. This worsens both adverse selection and deters borrowers from taking chances, increasing the possibility that a loan would fail. Unexpected changes in the overall inflation rate, its constituent portions, exchange rates, and interest rates will affect how viable potential borrowers are. The second impact of high inflation is that it makes it more difficult for banks to appraise loans. Additionally, asset prices under these circumstances will probably be quite volatile. As a result, it is also very speculative to predict the real value of loan security in the future (Badar & Yasmin Javid, 2013; Gizycki et al., 2001; Montiel & Servén, 2006; PricewaterhouseCoopers, 2019).

Liquidity support and prudential regulation

The deposit insurance policy instituted by the BoG, the primary reserve requirement (also termed as cash reserve ratio, CRR), does not adequately protect depositors from losing their deposits. The CRR is a minimum percentage of customers' deposits and notes that each commercial bank must keep in reserve rather than lending to customers. The required reserves are typically a percentage of cash physically held in the commercial banks' vaults (sometimes called vault cash) or deposits made with the BoG. The CRR typically ranges from 5% to 15%. The proportion of the deposits reserved will be paid back to depositors in the worst-case scenario where a bank goes bankrupt. An increase in the CRR denies the commercial banks the needed deposits for lending.

Conversely, the BoG utilizes the CRR to control the amount of cash in the economy to curb inflation. Local banks struggle when inflation is high, and the BoG raises the CRR to curb inflation. The situation is aggravated by the fact that most collapsed local banks had liquidity challenges and had not met the minimum capital requirement.

Additionally, the BoG's propensity to provide loans to the distressed banks rather than shut them down likely contributed significantly to moral hazard. The BoG, between June 2015 and November 2016, provided a total of GH¢620 million as liquidity support to Capital Bank instead of closing it down at that time is a case in point. The extent to which the failing banks were managed imprudently demonstrates grave regulatory and supervisory shortcomings.

Existing Practices and Policies to Mitigate Credit Risk

Credit risk management

Due to credit risk, banks must exercise judgment to maintain a reasonable distribution of liquidity among assets and accurately assess borrowers' default risks. The exercise of this judgement is termed credit risk management. Any bank's long-term profitability depends on effective credit risk management, which is a crucial part of a comprehensive risk management plan. To effectively manage credit risk, one must follow stringent credit criteria, diversify one's portfolio, fully comprehend the borrower's behaviours, and use precise monitoring and collection techniques. Selection, limitation, and diversification are the three guiding concepts for general credit risk management for loans.

Banks must exercise caution when deciding to whom to lend money. Credit decisions are made in accordance with the bank's delegation policies, and credit applications are reviewed by credit officers or credit committees. Limitation is the process by which banks establish distinct credit limits. The maximum sums that may be lent to particular individuals or organizations are expressly stated in limitation systems. Additionally, the proportion of large loans in total lending is restricted, and the size categories into which loans fall.

A maximum risk-to-total-assets ratio and a specific percentage of credit-risk-free assets, such as cash and government securities, are also requirements for banks. Diversifying credit management is necessary. Banks must diversify their clients, industry sectors, and geographic areas to avoid an excessive concentration of credit risk. As a result, large banks gain in this regard.

The bank's Credit Risk Management Policy determines the credit risk strategy because exposure to credit risk is still the main factor in bank failures. The target markets, risk acceptance/avoidance levels, risk tolerance limits, ideal degrees of diversification and concentration, credit risk assessment, monitoring, and regulating procedures are all spelt out in these policies.

There is evidence that the current active banks in Ghana practice (to a reasonable extent) good credit risk management (Apanga et al., 2016). This assertion is confirmed by the periodic sector reports by the BoG. For example, the BoG sector report of July 2021 stated that the financial soundness indicators of the existing twenty-three banks in the country remained strong, underpinned by improved solvency, liquidity, and profitability indicators.

Bank of Ghana legislations and directives

Existing legislation and directives to aid in sound credit risk management include the following.

1. The Banks and Specialized Deposit-Taking Institutions Act 2016, Act 930. This Act is the primary statute governing the banking industry in Ghana
2. The Ghana Deposit Protection Act 2016, Act 931. It calls for establishing a deposit protection programme, a deposit protection fund, the Ghana Deposit Protection Corporation (GDPC), and other related concerns. It aims to safeguard small depositors in the case of a bank failure
3. The Borrowers and Lenders Act 2008, Act 773, this Act establishes the legal basis for credit, as well as the rules for information disclosure needed by borrowers and lenders, among other things. It mandates that lenders notify the Collateral Registry of all fees and collaterals used by borrowers to obtain credit facilities from lenders.

4. The Credit Reporting Act 2007, Act 726. It addresses the legal framework for credit bureaus and establishes the prerequisites for credit reporting
5. The Credit Reporting Regulations 2020. This regulation contains an expanded list of institutions required to participate in the Credit Reporting System established under the Credit Reporting Act, 2007 (Act 726).
6. The Corporate Governance Directive, 2018. This directive is intended to help financial institutions adopt strong corporate governance standards, promote and maintain public trust, and reduce the likelihood of failures caused by weak corporate governance practices.

Operational Risk

This is the second most significant risk faced by banks. operational risk is the possibility that present technology, auditing, monitoring, and other support systems would fail or become inoperable. Operational risk emerges from flaws or deficiencies at either the technical (i.e., in a bank's information systems or risk management procedures) or organisational (i.e., in a bank's internal reporting, monitoring, and control systems). Technical operational hazards manifest themselves in various ways (such as errors in recording transactions, deficiencies in information systems or the absence of adequate tools for measuring risks). According to (Bessis, 2002), "the Basle Committee employs a widely accepted industrial definition of operational risk as the risk of direct or indirect loss due to insufficient or failing internal processes, people, or systems, or as a result of external occurrences".

Market Risk

Market risk is the possibility of financial loss brought on by unfavourable changes in the value of the trading portfolio due to movements in interest rates, stock prices, foreign exchange rates, or commodity prices. It happens when banks store stock as a kind of security or keep financial instruments on their trading books. The volume and scale of many large banks' trading portfolios have grown considerably, putting them at higher risk from the market. According to (Bessis, 2002), the idea of market risk includes the risk of loss throughout the time required to consummate a transaction (liquidation period). The two components of this risk are volatility and liquidity. In a volatile market, changes may be large even if the liquidation period is short. Second, selling some products in low-volume markets could be challenging (Bessis, 2002; Saunders & Cornett, 2008a).

Credit Scoring

As was noted in Section 1.4.1 above, one of the most crucial steps in banks' credit management decisions is credit risk analysis. This procedure involves gathering, examining, and categorising various credit factors and characteristics to evaluate credit judgments. The above-mentioned proven procedures fall short of solving all banks' credit risk problems. In order to reduce the current and future danger of a customer having bad credit, credit scoring is one of the essential techniques used to classify a bank's customers. Credit scoring's core idea is the classification of potential clients into high-quality applicants (those who can repay the loan) and low-quality applicants (those who, in one way or another, are unable to repay the loan). Due to improved data accessibility, more processing power, legal constraints, and demands for efficiency and economic growth, particularly in industrialised nations, the usage of credit scoring algorithms has considerably increased in recent years.

A borrower's creditworthiness is assessed using credit scoring, a statistical technique that incorporates a number of financial indicators into a single score. Consumer lending, credit card, and mortgage lending determine the chance of default or delinquency. Credit scoring calculates the likelihood that a prospective borrower, current borrower, or counterparty will default or fall behind on their payments. Banks use the credit score as a guide when deciding whether to grant credit. It is well known that credit scoring has a great deal of potential to support global economic growth. It is also useful for improving productivity, financial inclusion, and loan accessibility for private citizens and small, medium, and large organisations.

Any number range can be used to scale scores; generally, the lower the borrower's credit score, the greater the risk of a credit default. Risk-based pricing, a pricing technique in which the terms of a loan, including the interest rate paid to borrowers, are influenced by the borrower's credit risk, is used by banks and involves the use of credit scoring. Credit ratings are important because they affect a counterparty's ability to access credit, the conditions of credit facilities, such as the interest rates set by banks, and the choices made by potential investors.

In developed nations, banks frequently utilise credit scoring to approve relatively small loans. For loans up to US$100,000, which make up roughly 90% of all small business loans, most banks, according to Benchmark International, base their choice on the credit score (U.S. Small Business Administration, 2005). A highly developed proprietary scoring model has been created by banks worldwide using seven years' worth of internal customer data, including Ecuador's Banco Solidario. The bank is certain that it has gathered enough historical data to create a segmented, highly predictive database that will significantly boost its competitiveness. Additionally, Citibank's global lending model, known as Citi Business, acknowledges the limitations of a conventional custom scoring model and thus exclusively focuses on particular industry segments through a credit process that uses institutional knowledge about those industries to lend more wisely successfully.

The algorithms used to calculate credit scores have advanced recently. They have advanced from traditional statistical methods to cutting-edge technologies like artificial intelligence, which includes machine learning techniques like deep neural networks, random forests, and gradient boosting. In other cases, novel approaches have also increased the amount of data considered pertinent for the models and decisions of credit scoring. A study of 214 credit scoring studies (Abdou & Pointon, 2011) went into detail about the variables, the methods employed, and the performance rating criteria. See Table 1 below:

Table 1: Relative Predictive Accuracy of Different Classifiers using Credit Application Data

Author	Linear Regression	Logistic Regression	Decision Trees	Math Programming	Neural Nets	Generic Algorithms	Genetic Programming	K-nearest Neighbours	Support Vector Machines
Srinivisan and Kim (1987)	87.5	89.3	93.2	86.1					
Boyle et al. (1992)	77.5		75.0	74.7					
Henley (1995)	43.4	43.3	43.8						
Desai et al. (1997)	66.5	67.3			66.4				
Yobas et al. (2000)	68.4		62.3		62.0	64.7			
West (2000)	79.3	81.8	77.0		82.26			76.7	
Lee et al. (2002)	71.4	73.5			73.7				
Malhotra and Malhotra (2003)	69.3				72.0				
Baesens (2003)	79.3	79.3	77.0	79.0	79.4			78.2	79.7
Ong et al. (2005)	80.8		78.4		81.7		82.8		

(Abdou & Pointon, 2011; Crook et al., 2007)

The authors proved that there is not a single ideal statistical method for creating scoring models, and they confirmed that there is not a method that works in every situation.

Statement of the Problem

There has been an increase in the number of banks (both local and foreign) in Ghana over the past decade. These banks have made the banking industry in Ghana go through many structural changes. These adjustments include expanding the branch network, offering a broad range of financial services, and speeding up credit activity in various ways. The performance level of a bank is not measured based on only structural or physical growth but on its assets. The major assets created by banks are their credit facilities. These assets, i.e., the credit facilities, are created through the financial intermediation process. Laws requiring banks to invest in functional systems to track, monitor, and report on various transactions include the Banks and Specialized Deposit-Taking Institutions Act 2016, Act 930, the Credit Reporting Act 2007, Act 726, the Internal Capital Adequacy Assessment Process (ICAAP), IFRS 9, and FATCA. These are just a few examples of local and international laws. The limited use of credit scoring methods in most financial institutions in developing nations can be attributed to several factors. One issue is the lack of current, correct, and trustworthy data in credit bureaus and other data registries. Credit scoring is also hampered by the lack of access to credit reports and other pertinent financial information on business owners and their small firms. These elements limit the extent to which lenders may automate lending and lower the predictive power of the models.

Policymakers have made several noteworthy initiatives to enhance sector regulation, including the Bank of Ghana (BoG), the Ministry of Finance, and relevant governmental and financial industry regulatory bodies. On merits, the sector presently compares favourably to regional banking sectors regarding growth, profitability, and NPL ratio. Few banks, especially foreign ones, have a distinguished innovation, expansion, and value creation history. These achievements are reflected in their market value. However, the poor quality of the loan books that were prepared in previous years has resulted in an increase in default. For instance, Ghana's NPL ratio was 14.8% in December 2020 as opposed to 13.8% the year before (Bank of Ghana, 2021).

Thus, the failure of banks to deliver financially is due to the industry's high proportion of bad debt. A review of the sector contribution to the industry's NPLs showed that the top three sector credit recipients (Services, Commerce and Finance, and Manufacturing) constituted the top three contributors to the industry's NPLs. The services sector with the most significant credit allocation held a lower share of total NPLs of 9.2 % than the proportion of NPLs attributed to the commerce and finance sector of 18.5%. The manufacturing sector had a share of 17.2% in industry NPLs as of December 2020. The BoG encouraged banks to make major process and system changes in its sector report to modernise their stress testing, counterparty risk, and capital management infrastructure and improve performance.

The primary factor in the banking industry's earnings decline is risk management (credit, market and operational). The biggest type of risk facing financial organizations is credit risk. Borrowers have always had the option of breaking their promises for a number of different reasons, which exposes banks to credit risk. Losses arising from changes in portfolio value brought on by a real or perceived fall in credit quality just short of default could be included in these losses, as well as losses from complete defaults and losses from perceived credit quality declines.

The issue that this study aims to solve is the absence of a credit scoring tool used by the local commercial banks in their framework for managing credit risk. It is commonly known that credit scoring enables banks to accurately distinguish between loan defaulters and loan non-defaulters. The credit rating technology will improve credit risk management procedures already in place. Credit scoring will enable banks to recognize the various degrees of credit risk and how overall credit risk can be anticipated and managed, which will improve bank performance.

Banks must control the credit risk that comes with each credit and transaction and the overall portfolio. Additionally, the development of credit rules and processes in financial organisations is the responsibility of credit risk management. Bad debts and NPLs are globally the most obvious source of credit risk among financial institutions. In order to build a picture of a business's creditworthiness, Ghanaian banks regularly run into circumstances where they cannot link information about the firm and the business owner. This is a result of financial institutions' data becoming fragmented. Combining the more extensive data gathered during the loan application with information regarding later payments is difficult. Furthermore, even while this information is necessary for the development of high-quality credit scoring, information on rejected loans is not kept up to date.

Credit risk, as stated earlier, is the likelihood of suffering losses as a result of a debtor’s default. The banking sector has the strictest regulations worldwide. However, most local banks in Ghana do not employ scientific methods to measure and reduce credit risk. While the definition may be straightforward, measuring and reducing it are not. The use of mathematical and statistical models can alleviate the problem of measuring and reducing credit risk. Simply put, credit risk is measured by collecting data on industries, businesses and individuals who have contracted credit facilities and the frequency of default taken. The frequency of default is used to determine the probability of default (credit risk).

Only a few banks are keen on the proper allocation of funds for the specific type of credit to the right individuals or organizations in the right proportions to maximise returns on the disbursement of credit facilities. Credit facilities are usually given based on the type of credit the customer demands and the amount of money available. The other problem is that most banks in Ghana do not have discriminating models that differentiate between loan defaulters from loan non-defaulters. Given that the credit portfolio is a significant asset for banks, it requires continuous monitoring to support the growth of the bank. Therefore, conducting a study to help banks distinguish between loan defaulters and non-defaulters is crucial. This would enable banks to effectively manage their loan portfolios, minimize credit risk, and enhance their overall performance.

Ghana's banks can invest their capital in non-risky financial assets such as government bonds and treasury bills. However, such a move will derail the BoG's financial inclusion and economic stimuli agenda. Commercial banks are expected to grant many loans to customers to stimulate the economy and sustain themselves in the highly competitive market. This presents a problem for the banks. If the high percentage of non-performing loans in the bank's portfolio is not reduced, it could erode the capital basis of the institution and lower its profitability. The worst-case scenario is that the bank's inability to effectively manage its credit risk results in liquidation or insolvency. In order to appropriately monitor, manage, and minimise bank risks, it is crucial to understand what they are. The lack of a credit scoring model that can reliably distinguish between loan defaulters and loan non-defaulters affects bank decisions to limit or minimise credit risk. Diversification, lending to connected parties, overexposure, and making provision for loss or making allowances at a level sufficient to absorb the projected loss are all impacted by this deficit, as well as concentration and high exposures policies.

There is often just a small amount of empirical research on credit rating for developing nations (Kleimeier & Dinh, 2007; Vogelgesang, 2003). Researchers view risk management as a metric for assessing the success or failure of a financial organisation; nevertheless, models for credit scoring to support decisions have received little attention in Ghana. This research work highlights the usefulness of a credit scoring model to credit risk management.

Objective of the Study

The general objective of the study is to develop a credit scoring model for discriminating between loan defaulters and loan non-defaulters. This will help to minimise the credit risk of the selected Ghanaian commercial bank thereby maximizing the bank's profits.

The specific objectives are to:

1. Derive a logistic regression model to classify loan defaulters and loan non-defaulters.
2. Determine the optimal cut-off point for the logistic regression model and evaluate its performance against the other binary models that applied resampling techniques to handle imbalanced data.
3. Examine how the shifting of the cut-off point within the thresholds of the derived optimal cut-off point and the default cut-off point of 0.5 affect the credit risk and profit levels of the Ghanaian commercial bank.

Significance of the Study

Failures of banks are generally regarded as having a more significant negative impact on the economy and are therefore given greater importance than failures of other economic enterprises. A bank experiences an economic collapse when the market value of its assets is lower than the market value of its liabilities. Additionally, it is critical to limit loan defaults. The similar situation will arise if this is not done. Since both the lender and the borrower gain from the credit portfolio, systemic discrimination between loan defaulters and loan non-defaulters will essentially maximise returns.

Furthermore, the evolution in data availability and statistical techniques will offer increased possibilities and provide an extensive range of quantitative techniques, which can be used to develop models in credit facilities provided to clients of banking institutions. Additionally, the adoption of prediction models that distinguish between loan defaulters and non-debt defaulters of credit facilities may lessen the information imbalance currently present between Ghanaian banks and borrowers. The survival of banks and their clients depends on effective credit risk management. Let us say that lending carries a much lower risk. The banks will no longer be required to carry bad debts and use a portion of their revenues to settle them. Lower interest rates on loans and other kinds of credit will result from banks' increased interest in extending credit.

Limitations

This research was initiated to determine factors that contribute significantly to credit risk. These factors are to be determined from the KYC information gathered by banks at the account opening stage of customer onboarding; hence the data is secondary. The borrowing customers were limited to individual personal accounts because KYC data on corporate/business accounts are limited. The data on corporate/business accounts were on only registration from the Registrar General's Department, information on proof of identity and residence of the company's directors or business owner (in the case of sole proprietor businesses).

Additionally, this study is limited to only one indigenous Ghanaian bank. All banks guard their data protectively and refuse to grant academic requests, citing customer confidentiality. However, because all banks use the same minimum Know Your Customer (KYC) policy prescribed by the BOG, data from one bank is considered sufficient to represent the population under study.

Finally, there is incomplete data collected on borrowers by the bank resulting in missing data.

Organisation of the Study

This thesis is divided into five major chapters. The first discusses the general introduction, background information, problem statement, research aims, and constraints. Chapter two is devoted to a review of the literature, which includes a theoretical framework and empirical evaluations. Chapter three discusses the research methodology, the data sources, and the analytical approaches and methods employed in the work. Chapter four discusses the findings of the study. Chapter five is the concluding chapter; it contains a summary of the results and findings, recommendations, and conclusion.

Chapter Summary

The chapter introduces a study focused on Ghana's locally controlled commercial banks, prompted by the events of August 14, 2017, when two indigenous banks, UT Bank and Capital Bank, were liquidated as part of the Bank of Ghana's (BoG) initiative to clean up the financial system. The BoG's Asset Quality Review (AQR) program in 2016 revealed challenges in Ghana's banking sector, including lack of capital, increased non-performing loans (NPLs), weak liquidity and credit risk management, and poor corporate governance frameworks. The closure of several banks highlighted credit risk as a significant contributor to the collapse.

The causes of the problem are attributed to insolvency resulting from capital inadequacy, questionable transactions, weak corporate governance, and credit risk leading to high NPLs. Research supports the BoG's findings, linking serious banking challenges to lenient credit standards and poor lending practices by commercial banks, including insider lending to high-risk borrowers. Additionally, macroeconomic instability, lack of liquidity support, and prudential regulations exacerbate credit risk.

The chapter outlines the three primary roles of banks in a nation's economy: creating capital, assisting the central bank in maintaining currency value, and improving citizens' lives. Banks face various risks, with credit risk being the most significant in banking, as it involves the probability of loan defaults or loss of investments. Locally controlled banks in Ghana were found to engage in poor credit practices, leading to credit risk along with external factors.

The study's objective is to propose a statistical technique to mitigate credit risk, considering the challenges faced by Ghana's locally controlled commercial banks. The chapter sets the stage by providing essential background information, highlighting the problem's causes, and defining key terms related to credit risk and bank operations. It concludes by outlining the structure of the study, indicating the statement of the problem, research objectives, significance of the study, limitations, and the subsequent chapters' flow.

CHAPTER TWO

LITERATURE REVIEW

This chapter discusses the evolution of credit risk models used by financial organizations to minimize loss, from the pre-computer era to the present. The focus is on statistically valid approaches to modeling credit risk, as well as the potentials, shortfalls, and restrictions of the many methods used. By applying mathematical or statistical techniques, these academics aim to lower credit risk, particularly the chance of default.

Introduction

Most research works on credit risk analysis reviewed employed regression analysis to create credit scoring models and credit risk optimization models to assess and analyse risks and potential risks associated with loan portfolios. This review is a treasure trove of critique on relevant articles and research works conducted on credit risk management and measurement.

Regression Analysis Models

Most recently, research on "Risk Management and Performance of Listed Banks in Ghana" by (Ofosu-Hene & Amoh, 2016) used regression computations to analyse credit risks. The research sought to establish an overall risk index to determine the risk level of banks listed on the "Ghana Stock Exchange (GSE)" and further determine whether a salient link exists between bank performance and risk management. The research recognizes that a vast number of works in the literature have examined the risks and performances of banks. However, majority of these studies focused on the entire banking industry. So, the research problem addressed here is to fill the gap left by limited evidence that show risk index of listed banks on the GSE and to assess how the risk index impacts their performance in Ghana.

The abstract of their study makes it obvious that creating an index for the listed banks is not a straightforward task. The authors began the study by outlining three key justifications for managers' involvement in activities which ensures risk management. The first is the managers' personal interest in maintaining their position and fortune within the company. The expense of potential financial hardship is the second justification. Losses of returns can result in stakeholders disinterest and doubts in the financial organization's operations, a severe decline in the industry's competitive tactical status, the revocation of a licence, or even insolvency. The capital market's imperfections, one of the causes of risk management, are the last justification (Oldfield & Santomero, 1997). The authors thought that one strategy to address the complex problem of why managers engage in risk management activities was to increase stakeholder confidence by going public. The article review by (Ofosu-Hene and Amoh, 2016) summarises the easily available literature on the key risks affecting banks, advice on managing these risks, and empirical literature.

The probability that counterparties to a loan or derivative transaction would default is referred to as credit risk by (Koch & Scott, 2014) when discussing the main risks affecting banks. This is when a participant in a deal does not uphold their end of the bargain by paying the interest and principal at the predetermined period. The authors also acknowledge that this explanation of credit risk aligns with the view of (Fabozzi et al., 2003). Literature from (Saunders & Cornett, 2008) is also used to adequately explain firm-specific and systematic credit risk. Moving on to market risk facing banks, literature from the authors describe the market risk concept. The explanation for liquidity risk is ably supported by (Gup & Kolari, 2005). Also, (Kanchu & Kumar, 2013; van Greuning & Bratanovic, 2009) are cited as sources of literature for an explanation of interest rate risk.

A relevant literature review provided on the issue of management of risks with (Cebenoyan & Strahan, 2004; Moreno, 2006; Turner, 2006), all cited as sources to explain the relevance of executing a risk management program. For the empirical literature review, the portfolio theory of (Donaldson, 2000) is explained to give an insight into the new model for performance-driven organisational change where risk plays a central role. Sources such as (Adedapo Soyemi et al., 2014; Bloom & Milkovich, 1998; Nocco & Stulz, 2006; Schroeck, 2002; Tandelilin et al., 2007) are cited to throw more light on risk management and financial performance, where further evidence from (Bettis & Howard, 1990; Carey & Stulz, 2005; Merton & Perold, 1993; Smithson & Simkins, 2005) explains the effects of lower risks on bank performance. Further, (Boahene et al., 2012; Odonkor et al., 2011) are cited to explain that lower risk levels affect bank performance in the Ghanaian context.

The article's methodology was based on information from seven selected banks that were listed on the "Ghana Stock Exchange (GSE)" between 2007 and 2014. Additionally used secondary data came from the banks' audited yearly accounts and the Bank of Ghana's statutory returns. Additionally, data from the databases of the "Ghana Statistical Service" and the "Bank of Ghana", respectively, were obtained to provide information on inflation and the exchange rate. The performance metrics of return on assets (ROA) and return on equity (ROE) were chosen as a consequence of using the method of (Naimy, 2005) to select the performance indicators. In addition, the accounting profit per dollar of equity capital's book value was determined. Divide net income by total equity to arrive at this measure. Since total assets are calculated as total assets divided by total equity because net income is defined as total assets divided by net income. This provides a measurement of the base equity multiplier for leverage at an institution.

The Risk Index established by (Hannan & Hanweck, 1988), which gauges perceived bank insolvency, was chosen as the risk index for the study. This risk indicator is generated by combining CAP, ROA, and ROA's standard deviation. In addition, (Dietrich & Wanzenried, 2011), similar to that employed by (Odonkor et al., 2011; Sakyi et al., 2014), was used to evaluate the number of variables that affect the performance of financial institutions. This modified panel regression model is how it manifests itself empirically.

Using the methodology described in the previous sentence, it was found that the risk index for individual banks listed on the Ghana Stock Exchange (GSE) increased from a mean of 25.93 in 2007 to a climax of about 33.3 in 2010 and then gradually decreased to about 26.54 in 2013 before experiencing a sharp decline in 2014 to reach 8.58. Additionally, there was no discernible connection between risk management (RI) and the bank's performance according to the panel linear multiple regressions that were utilised, which used the Least Square Estimator for ease of explanation. According to the authors, the findings are at odds with the research from (Jafari et al., 2011), which suggests that the performance of businesses can increase by controlling their exposure to risk factors. As a result, a crucial beneficial association exist between management of risk and business performance.

In a related study, (Jonathan, 2012) used logistic regression analysis to identify high-risk individuals and, if possible, make conclusions about credit risk. The study looked at the creditworthiness of Atwima Kwanwoma residents, specifically the Atwima Kwanwoma rural bank. Although credit risk is a global concern, it is essential to remember that a person's environment, or geographic location, tends to affect their risk probability. Credit businesses frequently employ scoring models that categorise consumers as either "excellent" or "bad." The scoring method has a flaw in that it ignores the state of the economy, as well as the social and political context. According to Jonathan, credit risk cannot exist independently of financial risk because these two risks are interdependent. The same source that puts the institution's credit at risk could also put it at risk for other things. For instance, a poor portfolio could draw liquidity issues.

In the introduction to his article, he provided a brief overview of Ghana's banking system, including the history of the country's first bank, the "Bank of British West Africa (now Standard Chartered Bank)", which was founded in 1896, and the developments that led to the emergence of rural banks in Ghana. A brief history of "Atwima Kwanwoma Rural Bank" was also provided to emphasise the significance of rural banking in Ghana.

In order to identify the creditworthiness of borrowers based on a number of observable borrower characteristics, determine the model's predictive power and the implications for use of the model in the rural banking sector, and make policy recommendations centred on the thesis' findings, he took a sample and grouped the data by gender, marital status, dependency ratio, occupation, amount of loan collected, and age. To establish the association between the variables, he next applied the logistic regression model. The first round of data collecting was done by the Atwima Kwanwoma Rural Bank in Pakyi Kumasi, Ghana's Ashanti Region. He said that the data was secondary and comprised demographic and behavioural characteristics of present (within the last five years) and potential bank customers. For computations and analysis, SPSS and Excel were utilized. Using logistic regression as a method, the credit risk model was developed. Although only a small part of the bank's whole clientele was represented in the sample, the results were encouraging. 270 women and 330 males out of the bank's 600 loan recipients were considered. In comparison to women, men had a higher default rate. Compared to single borrowers, married borrowers had higher default rates. With the rise in the default rate came an increase of dependents. Self-employed individuals recorded higher default rate than individuals hired by an employer. To assess a borrower's creditworthiness, credit risk models (logistic regression models) can be easily created using the logistic regression approach.

Based on his analysis, he then offered some advice. He suggested that the bank engage qualified personnel to analyse credit risk; he also mentioned that they could look at all the client's information and make an important determination about a client's creditworthiness. Additionally, he suggested that the bank occasionally provides entrepreneurship workshops for clients, particularly self-employed ones. In order to prevent the loan from being used for something other than what it was intended for, consumers will receive instructions on how to use their loan. Additionally, the bank can put up a comprehensive database of all personal loan situations. He said that the use of the database to modify the upcoming policy of personal consumer loans, and the extension of credit tactics are enhanced. Also, the outcome can serve as a model for developing relatively new financial goods. The bank is free to continue looking for distinctive elements affecting consumer loan applications' credit risk.

It is possible that the social and demographic factors influenced his decision. He advocated, for example, that a person's gender be considered when determining their creditworthiness, but in the modern world, this does not form a basis for equality. Credit denial based only on a borrower's gender would be discriminatory. Certain social factors may have influenced the high-risk status of the guys in Atwima Kwanwoma, but this does not necessarily mean that all males are in this situation. Additionally, he limited his independent variables. In addition, other independent variables such as "level of education", "number of years with current employer", "number of years at current address", and "household income", could also be used to estimate a borrower's trustworthiness "(Sex, Age, Occupation, Marital Status, Number of Dependents, Amount of Loan Collected)". Overall, Paddy Jonathan presented some useful solutions and shed some light on this problem that is growing more and more pervasive in our culture.

The following objectives were addressed in a similar research study by Poudel (2012), which used descriptive statistics and regression models to examine "The impact of credit risk management on the financial performance of commercial banks in Nepal.” The objectives were to determine the extent to which rate of default, debt collection, and cost per loan asset influence the banks’ performance. This study aimed to close gaps in prior research and offer the first comprehensive examination of efficient credit risk management. The research work’s abstract clarified that the researcher looked at various characteristics related to credit risk management related to banks’ financial performance. Three features of this study were the “capital adequacy ratio”, “cost per loan asset”, and “default ratio”. The study used both correlational and regression analysis to analyse the collected data. According to the study, all of these factors negatively impact banks’ financial health; nonetheless, the default rate is the strongest predictor of health. The study advised banks to develop and implement plans to reduce their exposure to credit risk and boost long-term profitability.

The researcher stated before that he intended to conduct extensive research on the influence of credit risk on the financial performances of banks. The researcher recognised contributions of other researchers and used some applicable knowledge in his work. According to the study, insufficient credit policies continue to be the primary cause of major financial issues in the banking sector. Due to this background, effective and efficient credit risk management has come under more attention in recent years. According to the study, a bank’s risk-adjusted return rate must be maximised through the prudent management of credit exposure within predetermined limitations if credit risk management is effective. Additionally, the overall portfolio credit risk and the risk associated with specific credit transactions must be effectively managed by the banks. According to (Bernanke, 1993) whose work the researcher read, the process of creating credit runs smoothly when there is transfer of money from ultimate savers to borrowers. The researcher's study of the works of (Koford & Tschoegln, 1999) revealed that banks regularly experience bad lending practices. Credit risk is known as the possibility that a bank loan will not achieve full or partial return rate in due time (Campbell, 2007). According to Campbell's writings that the researcher has read, there were numerous potential risk sources. These included political, market, liquidity, and credit risks, as well as interest rate, market, and foreign exchange risks. However, the principal risk banking and financial institutions face is credit risk (Gray, Cassidy, & RBA., 1997). The researcher also looked into the writings of (Bryant, 1999), which showed that before the financial sector's deregulation, banks were strongly urged to offer lending facilities to customers and potential customers who could easily express their creditworthiness. Before authorising or awarding new loans or extending existing credit, banks must establish a clear methodology as part of their credit risk management procedures.

As well as scrupulously adhering to monitoring regulations, these procedures also take other significant steps to minimize or completely eliminate the risks associated with related lending (The Basel Committee, 2000). (Boyd, 1993) found that analysing loan applications and preserving the integrity of the bank's overall loan portfolio relies on the processes and controls used for granting credit. The creation of a secure environment minimising credit risk, sound credit granting practices, proper credit administration, measurement, monitoring, and control of credit risk, policies and strategies that concisely outline the scope and allocation of bank credit facilities, and the way a credit portfolio is managed, that is, how loans are made, valued, monitored, and collected, are all essential elements of effective credit risk management (The Basel Committee, 2000). The investigation conducted by the researcher also revealed that private banks use effective credit risk management techniques with greater care than state-owned banks.

In comparison to other studies on the subject under examination, the methods and data type used for this study were not as distinctive. The study population includes all banks functioning in Nepal, and 31 banks were employed in the research. The study spanned the years 2001 to 2011 due to the numerous transformations the banking sector witnessed during this time. In order to analyze the gathered data and the descriptive research design employed for the study, the researcher additionally used regression analysis and correlation. The research used secondary data. To analyze the data, trend analysis was used to compare the profitability ratio to the “default rate”, “cost per loan asset”, and “capital adequacy ratio”. During the research period, the profitability for each year was calculated. Regression statistics were used to further evaluate the ratio using the SPSS program version 20.

The researcher defined the crucial parameters stated below in order to include them in the model specification; Return on assets is the proportion of a company’s profits before interest and taxes (EBIT) to its total net assets (ROA). The ratio is thought of as a gauge for how successfully and efficiently a corporation is utilising its resources prior to having to fulfill contractual obligations. It is determined as follows:

“Default rate (DR) is the term for practice in the financial services industry for a special lender to alter the terms of a credit facility from the normal terms to the default terms, that is, the terms and rates given to those who have missed credit payments (Appa, 1996). DR ratio can be computed as; $DR\ Ratio = Non - \frac{Performing\ Loans}{Total\ loan}$,,.

"Cost per loan asset (CLA) is the mean cost per loan paid out to a customer in monetary terms. This purpose indicates efficiency in disbursing loanable funds to customers (Appa, 1996). CLA ratio can be calculated as; $CLA\ Ratio = \frac{Total\ Operating\ Cost}{Total\ amount\ of\ loans}$. Capital Adequacy Ratio (CAR) is a measure of the amount of bank's capital expressed as a percentage of its risk weighted credit exposure. CAR can be calculated:"$CAR = \frac{Capital\ fund}{Risk\ Weighted\ Assets}$.

The researcher specified this model:

Performance (ROA) $= \beta_0 + \beta_1\mathrm{DR} + \beta_2\mathrm{CLA} + \beta_3\mathrm{CAR} + e_{it}$ from this econometric model $Y = \beta_0 + \beta F_{it} + e_{it}$

where Y is the dependent variable. β_0 is constant, β is the coefficient of the explanatory variable, F_{it} is the explanatory variable and e_{it} is the error term (assumed to have zero mean and independent across time period).

The capital adequacy ratio, one of the risk management indicators, comes in second with a 25% ranking, while the default rate, another indicator, is a good predictor of bank financial performance with a 56% ranking. Cost per loan asset is not a useful predictor of bank performance, making default rate management the only risk management indicator with the greatest ability to predict bank performance. The research work showed that credit risk management is crucial to bank performance since it has a considerable impact on bank performance and can account for up to 22.6% of it. The researcher also suggested that more study be done in order to manage credit risk properly and enhance bank financial performance because other factors that were not investigated in this research had a considerable (76.4%) impact on bank performance.

The article "Credit Risk Management in Commercial Banks" by (Konovalova et al., 2016) proposes a model for credit risk assessment based on factor analysis of retail clients and borrowers in order to ensure predictive management of the level of risk faced by potential clients in commercial banks engaged in consumer lending. In addition to doing a quantitative assessment and analysis of the credit risk and rating of borrowers, managing credit risk is a crucial concern for all banks involved in lending to individuals and businesses. The study's major goal is to assess the degree of risk posed by various groups of retail borrowers in order to reduce and prevent credit risk in the future and improve the management of banking risks.

It is clear from the beginning of the article that managing credit risk is one of the most important tasks for preserving financial stability and liquidity. As a result, the authors move on from the theoretical underpinnings of credit risk management to the following quotation (Kiselakova et al., 2015; Kiselakova & Kiselak, 2013). According to (Korobova, 2010), credit risk assessment of the borrower entails researching and evaluating both qualitative and quantitative indications of the borrower's financial situation. Additional details on evaluating risk variables are provided by (Rodina et al., 2013) using relevant literature, such as (Dmitriadi, 2010; Andrianova & Barannikov, 2013). Additionally, the assertion that loan classification is unstable is supported by (Solojentsev, 2004). The authors of (Seitz & Stickel, 2002) use the problem of instability in loan classification to emphasize the necessity of credit risk analysis and quantification procedures for a bank's effective operation. To measure and analyse credit risk, each bank develops its own risk probability assessment model, incorporating the broad suggestions of the Basel Committee on Banking Supervision. Therefore, the only way to create a reliable risk assessment model and properly manage credit risk is by ongoing quantitative analysis of statistical data on credit success. The article also emphasizes that while there are many ways to calculate the credit risk that a certain borrower faces, systems that are based on mathematical models are more reliable and successful than those that are not, according to experience from around the globe.

According to the article's discussion of the credit risk assessment model, in order to guarantee effective and efficient credit risk management in commercial banks, it is critical to decide what terms and conditions should apply to bank customers who take loans in a way that both appeals to potential borrowers and ensures loan repayment. This is accomplished in the essay using clustering as a type of credit risk assessment. Additionally, the authors of this article ran model experiments using statistical data on the credit histories of clients of Latvian commercial banks involved in consumer lending. Both Excel and the statistical program SPSS Statistics were used to process the data. The model sample contains data on variables such as the loan duration (month), the loan amount (value), the borrower's sex (0 - woman, 1 - man), the borrower's age (age), the borrower's number of children (number of children), and the borrower's average earnings (income). Each borrower also receives a variable condition that describes the possibility of economic harm and whether or not there will be problems making loan payments (0 means there won't be any problems, 1 means there are)". This investigation is completed in phases. Stage one was defining the initial conditions of the problem. Finding each factor's eigenvalues was the second step. Studying factor loadings is stage three. The fourth stage specified the number of relevant factors using the "Varimax row" method. Stage five assessed the effectiveness of the discovered solution and produced a correlation matrix to validate the precision of the pertinent factor selection. The article gives a thorough explanation of how each stage's results work.

In order to calculate the credit risk for each class, the following additional variables were also calculated: 1) the anticipated outcome of a credit transaction; 2) the range of potential outcomes of the operation relative to the anticipated value (result); 3) the dispersion and the mean linear deviation; and 4) the credit risk based on the mean linear deviation and of the operation's most anticipated outcome (cluster). A differentiated approach to credit risk management was also made possible by the article's authors' use of the credit risk assessment model that was advanced. Thus, credit risk management was defined as a procedure involving the following steps: identifying risk factors; assessing the potential effects of those factors; choosing managerial strategies intended to lessen the effects of a particular risk factor; and observing (monitoring) the execution of the chosen strategies intended to lessen and neutralize the impact of a particular risk factor The article goes into great detail to explain each step. The article goes on to say that credit risk control is essential and that the core principle of credit risk management is the integration of all the stages of credit risk management.

The loan amount, the loan term, and the borrower's typical income, according to the article, are the three factors that have the greatest impact on how much of a credit risk a bank assumes when lending to individuals (retail clients). The article also claims that the primary source of information used to estimate the credit risk level for banks that extend credit is the client's credit history. The paper goes on to say that the factor analysis method outlined in the study should be used to estimate credit risk in commercial banks. Furthermore, the core of this suggested methodology should be the application of clustering, dispersion analysis, and primary component analyses techniques. The research's future focus will be on "credit risk management" in circumstances where loans are given to other borrower types, like small and medium-sized businesses (SMEs) and large industrial businesses. The study's scope was constrained because it only examined loans given to retail clients. This limitation will determine how wide the research can be expanded. However, the article's authors do alert readers to the study's limitations.

By utilizing both a logistic regression analysis and a multiple discriminant analysis, the author of the article "Assessing Credit Default Using Logistic Regression and Multiple Discriminant Analysis" (Memic, 2015) set out to determine the likelihood of credit default on the banking sectors in Bosnia and Herzegovina. To increase the precision of default predictions, a multiple discriminant analysis and a logistic regression analysis are required. These problems are comparable to those that the banking sector encounters when deciding which group of individuals or institutions can be regarded as viable or not viable for credit consideration.

The researcher discovered that the increased rates of default occurrence are due to the resurgence of the global financial crisis, which is seen by most developing and developed economies as the worst financial disaster since the Great Depression. The Basel Committee on Banking Supervision states that a default would be deemed to have occurred in any of the following circumstances: (1) the debtor is unlikely to pay its debt obligation; (2) the debtor is past due more than 90 days on any credit obligation; (3) the debtor has filed for bankruptcy; or (4) a credit lost event. The different methods of default prediction were assessed using the time to default.

The main inquiry of the study was: “can credit default in Bosnia and Herzegovina be predicted using traditional logistic regression?”; “is traditional logistic regression useful for predicting Bosnia and Herzegovina’s credit default?”; “are all business sizes and locations in Bosnia and Herzegovina expected to experience credit default in the same way?”.

Both Bosnia and Herzegovina’s banking institutions provided information for the analysis. Multiple discriminant analysis and logistic regression models were employed by the researcher. The study assessed financial ratios to distinguish between defaulted and sound businesses in the selected sample. Companies that went out of business were named along with the pertinent years that they occurred. When scrutinizing the logistic regression model, it can be seen that returns on the assets under consideration were statistically significant during each of the four possible default periods. It was discovered that credit default prediction was distinct from multiple discriminant analysis and logistic regression. When compared to multiple discriminant analysis, it was found that logistic regression had a higher capacity for prediction. The difference in default prediction rates between Bosnia and Herzegovina's financial economies, however, was not statistically significant.

Optimisation Models

Most research works employ the use of regression models to analyse credit risks; however, the research work on "Credit Risk Optimisation with Conditional Value-at-Risk Criterion" by (Andersson et al., 2000) was determined to explore a new approach to credit risk optimisation, an approach which was proved to have been an improvement on the existing methods of credit risk optimization.

Based on a "Conditional Value-at-Risk (CVaR)" metric, the novel model discussed in this article assumed that the projected loss would be greater than the Value-at-Risk (VaR). According to the abstract, the credit risk distribution is produced by Monte Carlo simulations, and the optimisation problem is successfully solved by linear programming. Other names for Conditional Value-at-Risk (CVaR) in the article include Mean Excess, Mean Shortfall, and Tail VaR. According to the paper, the algorithm used was quite effective; it could handle thousands of scenarios and hundreds of instruments in a fair amount of computing time. An emerging market bond portfolio is used to illustrate the method.

This paper seeks to propose and assess a new approach for minimizing credit risk portfolios. The article revealed credit risk management to be the core activity of most financial institutions. Unlike the new approach, which believes that mean and standard deviation are the two statistical measures that can be used to balance portfolio return loss as employed in this article, the primitive methods assume that portfolio return loss is normally distributed, hence the Pareto optimal point was used to determine the portfolio with the best mean-variance point.

Some observable limitations associated with the Value-at-Risk (VaR) was that credit losses are characterized by large likelihood of small earnings coupled small chance of losing an important investment. The loss distribution was revealed to be heavily skewed; this led to inadequacies in market risk standard optimization tools. The lack of historical data poses significant challenges in risk modelling, making it tedious to estimate credit correlation. Mathematically, VaR is seen as a non-smooth, nonconvex with respect to position, lack of sub-additivity and multiextremum functions, making it difficult to control and optimize risk.

By contrast, CVaR was preferred over VaR because it is convex with respect to position. CVaR is mathematically greater than equal to VaR, where a portfolio with CVaR might also have low VaR. In order to address the skewed return loss associated with VaR, the researchers adopted CVaR as the preferred risk measure. Although the VaR was useful at predicting, it failed to provide adequate information on excess loss, which may be significantly greater.

Approaches to the minimization of "CVaR were described. (Bucay & Rosen, 2001) was reviewed to have applied credit matrix methodology to a portfolio of corporate and sovereign bonds issued in emerging markets. Credit Matrices is a tool for assessing portfolio risk due to defaults and changes in the obligor's credit quality, such as updates in credit ratings. The recovery rates are presumed to be constant and equal 30% of the risk-free value for all obligors. Mausser and Rosen were reviewed to have conducted scenario optimization of the portfolio with the expected regret function. A vivid description was given of the portfolio; the portfolio consists of 197 emerging market bonds issued by 86 obligors in 29 countries. The portfolios that were examined had a mark-to-market value of 8.8 billion USD. As explained in the article, the Monte Carlo simulations examined the one-year portfolio credit loss distribution, and linear programming was employed to deal with larger data sets".

After gathering enough viable data for the research, the researchers proceeded to conduct relevant analysis to arrive at the desired conclusion. The researchers considered the following steps: optimal hedging, a re-balancing of the portfolios was considered by changing the position of a single obligor while holding other relevant positions constant; it was revealed that portfolio credit risk could be minimized by minimizing CVaR in order to obtain the optimal contract size. Minimise CVaR, all positions under consideration were optimized, and the linear programming problem was solved. Relevant positions were bounded to prevent long and short positions in any obligor. Risk return was also considered by reviewing different performance functions.

After taking into account the offered data in the context of the procedures used to arrive at the anticipated outcomes, the researchers were able to draw a statistical inference that was largely relevant. The usage of the CVaR does not require conducting sensitivity research in regard to the regret threshold, unlike the Maximum Expected Approach. In order to automatically determine the required threshold, that is, the VaR that corresponds to a similar confidence level, the research revealed the minimal CVaR approved.

In this article, banks and other financial institutions were cautioned to have sufficient reserves to cover anticipated losses and capital to cover unforeseen losses. Theoretically defined as VaR minus the expected loss, the highest loss at a given quantile was used to explain the unexpected loss. This study's novel credit risk optimisation technique has produced a comparable portfolio return to prior rudimentary methods while reducing predicted losses by over 100% and unexpected losses by almost 80%.

Correlation Analysis

By utilising descriptive analysis and correlation regression, the paper on Credit Risk Management Practices of Commercial Banks in Kenya by (Afande, 2014) intended to evaluate the present practises of credit risk management by commercial banks in Kenya. The purpose of the study was to evaluate the factors that affect the extent to which credit risk management practises employed by commercial banks are effective. In addition, the study sought to assess how Banks in Kenya use these credit management practises in their daily operations and mechanisms for dealing with various types of risk, and to look at the internal performance measures of bank lending used by commercial banks.

As indicatedin the article's introduction, credit risk is a major problem that Kenya's commercial banks are currently facing. Although it emphasizes that credit risk poses the biggest threat to banks' bottom lines, it divides risk into three main categories: financial risk, operational risk, and strategic risk. It is significant to highlight that Kenyan bank, both old and new, have experienced significant problems, but credit risk repeatedly emerges as a major root cause of banking challenges. However, in the majority of transition economies, particularly Kenya, loan activities have proven contentious and difficult. This is due to business groups complaining about a lack of credit and banks' unreasonablely high standards, despite the fact that commercial banks have suffered large losses as a result of bad loans.In order to minimize credit losses and associated credit risk, commercial banks must have a successful credit risk management system in place. Despite all of these challenges, bank lending remains a significant source of revenue and is essential for success.

Due to the necessity of bank lending in revenue generation and overall success of banking institutions, effective credit risk management strategies are required.According to (Ziegel, 2001), which is referenced in the article, prudent lending practices include three crucial elements; the systematic identification of each loan applicant's risk, the modification of lending conditions prior to loan approval to account for this risk, and prompt execution of arrears processes when payments are delayed or refused. There are many articles on managing credit risk generally, but very few that target the problem in Kenya, necessitating the writing of this study. The study's main objective was to investigate the methods Kenyan commercial banks employ to control credit risk. The article also attempted to assess the degree to which commercial banks use internal lending performance metrics and credit risk management practises and procedures to manage various types of risk, as well as the factors that influence how effectively these practises are implemented by commercial banks.

In its literature analysis, the paper defines credit risk as the likelihood that a bank debtor or counterparty will not fulfil their commitments in accordance with the terms set forth. The article discusses credit risk in financial instruments other than loans, such as acceptances, interbank transactions, trade financing, foreign exchange transactions, financial futures, swaps, options, bonds, and stocks, together with the granting of commitments or assurances and transaction completion. The work by (Oldfield & Santomero, 2000), which provided the steps to risk management approaches, was cited. The four steps recommended for good risk management were the development of standards and reports, the enforcement of position limitations and laws, the development of self-investment regulations and techniques, and the co-ordination of incentive contracts and remuneration. Wyman also contends that while liquidity risk and the mix of loan products are insensitive to profit efficiency, credit risk and insolvency risk are very sensitive to profit gains. The four primary categories of risks were established using (Jeremy & Stein, 2000). These categories included credit risk, interest rate risk, market risk, market liquidity risk, and systematic risk. The article makes a second argument that there are two primary types of risk. These two types of risk are systematic and unsystematic. He says that systemic risk has to do with the economy or the market as a whole, but unsystematic risk has to do with a particular asset or company.

A descriptive research methodology was used to carry out the study. The population was made up entirely of Kenya's commercial banks. The sampling frame, which was taken from the Central Bank of Kenya, consisted of 45 items. The stratified random selection method was used to pick 33 respondents, representing 70% of the population. For these banks, Tier 1, Tier II, and Tier III groups were developed. A semi-structured questionnaire was used to collect the data from the banks. Data referring to the background of the respondents was analyzed using content analysis, whilst data pertaining to the research's goals were examined using descriptive statistics including average, median, and mode. Studies on correlation were done. The data were displayed using frequency tables, charts, and bar graphs. Correlations, a statistical method that may establish if and how strongly two variables are related, were used to assess the effectiveness of credit risk management in relation to internal performance measures.

According to the report, commercial banks in Kenya use credit risk management strategies such as careful loan analysis, the pursuit of collateral, and credit histories of borrowers. As defensive risk management techniques, the bankers also use agreements, credit restrictions, loan securitization, and loan syndication. The development of a credit policy which precisely defines the extent and distribution of bank credit facilities, the maintenance of a credit administration system with adequate credit controls, top management support, communication of credit guidelines and policy to stakeholders in the credit department, screening of prospective borrowers, and the recruitment of qualified staff are among the factors that impact the efficiency of credit risk management systems used by co-operative banks.The internal performance measures of bank lending used by commercial banks in Kenya include Basel II criteria and bank profitability, including return on equity, return on assets, and return on investment. The benchmarks that have been established also include directories for the cost of each loan that has been completed, the cost of loans per $1,000, the noninterest income from loans, and the number of loans per employee.

Other Credit Scoring Models

(Harris, 2013)'s study; "Default Definition Selection for Credit Scoring," sought to determine the effectiveness and precision of credit-scoring models used to evaluate customer credit risk, discuss the major contributing factors to the recent financial crisis, and propose an advanced algorithm for default definition sampling in the credit scoring research.As stated in the abstract, the optimal default definition selection (ODDS) is recommended to enhance credit-scoring for credit risk assessment, as opposed to prior to this when credit risk was assessed using subjective methods based on staff expertise and judgement. The algorithm was used to choose the default definition for the random forest tree algorithm to evaluate the ODDS. The categorization models that resulted were compared to those that were created using the default definition, which was not used.According to the final results, models developed using the ODDS algorithm's default definition choices were statistically better than models created using other definition choices.

The need to develop more accurate credit risk assessment techniques has become a matter of concern to heads of financial institutions across the globe. Most mortgage-providing institutions experienced a financial downturn due to the collapse of the securities that back these facilities. Some of the causes associated with this financial crisis were the (1) granting of consumer credit and (2) providing corporate loans. The more advanced ways of assessing customer credit risk have proven to be useful to managers of financial institutions other than the primitive ways of assessing credit risk.

These methods, as applied to customer and potential customer credit risk assessment, include determining whether a credit applicant is likely to default or not based on statistical methodologies obtained from past customers' socio-demographic data, such as employment status, revenue, and expenditure.

A default is deemed to have occurred when: (1) the financial institution believes the debtor is unlikely to return its entire debt and fails to sell the securities it holds to pay off the remaining debt; and (2) any major sum owed to the specified financial institution by the principal debtor in at least 90 days.The researcher then suggested using other default times, such as 30 or 60 days past due dates, with the help of a classification method for vector machines. The study found that default definition significantly affects the performance of the specified model. The researcher identified some of the contributing factors to the financial crisis.The financial laws passed by the Congress of the United States of America (USA) on the Equal Credit Opportunity Act (ECOA), the Fair and Accurate Credit Transaction Act (FACTA), the Fair Housing Act (FHA) and the Community Reinvestment Act (CRA) were purposed to regulate the granting of credit and mortgage facilities. These laws allowed the financial institutions to advance credit facilities to customers who were ascribed as potentially high risked customers but acceptable for loan facilities. As a result of these financial practises, financial institutions were: (1) exposed to a wide range of defaults resulting from their lending activities; and (2) prevented the institution from advancing new loan facilities using sunk capital needed by regulators to cover possible default risk. According to the study, these fair financial values were pro-cyclical because the US housing market slump and related asset markdown losses caused the capital base to deteriorate, lowering the capital institution’s creditworthiness rating and raising the cost of capital as a result. These and similar factors formed the basis of the financial crisis.

Barbados, a Small Island Developing State (SIDS), was hit financially by the global financial crisis. The Central Bank of Barbados reported in a Financial Stability Report that local banking firms had increased delinquency rates since the 2007/2008 fiscal year. It was indicated that the Barbados financial institutions were slow at adapting to the changing trends in credit risk assessment which implied that most of the financial institutions adhere to the primitive ways of credit risk assessment.

Discriminant analysis, as suggested by Fisher in his foundational work, was to help in discriminating between two or more populations. The researcher shed further light on the credit scoring processes. As a tool, credit scoring has been useful in determining a credit applicant's creditworthiness. Credit scoring is said to be automated, operates in a condensed processing time for evaluating potential consumers, and derives its conclusions from defined statistical models that use data rather than human judgments that are subject to human subjectivity.

The researcher made an advantage of a financial institution’s dataset from Barbados. Before cleansing the information, this dataset examined 18 client attributes over 47,407 entries from 2007 to 2010. The variables included the applicant’s marital status, the period of stay at their present address, the number of dependents they had and the ages of up to eight of the dependents In addition, the employment status of the applicant, the number of years the applicant has worked for their current employer, the requested loan amount, the type of loan and its purpose, their monthly income, and their age were some of the demographic factors included in the dataset.

The obtained sample A was under-sampled to create a new sample B. For analysis, the sample was divided into three parts: test (20%), cross-validation (20%), and training (60%). In addition, the researcher divided the training program into three additional groups and assigned different default definitions to each. The models were built using cross-validation and training datasets. The cross-validation data file was used to evaluate the training models that were created. The ultimate classifier was built using the training data set with the highest performance. The dataset was also analysed using the Analysis of Variance (ANOVA) test and the Bonferroni post hoc test. To create credit rating models for Barbados' financial institutions, the researcher used a random tree. The random forest method creates a strong classifier by combining a number of weak decision trees. According to the study, 60-day past default definition models were statistically superior to 30 and 90-day models. The existing works are thus improved by using the ODDS algorithm to choose the best default definition from a given search space. Future work by this author will aim to address and enhance the ODDS algorithm and take various methods of choosing the best default definition into consideration. Future research will also look at how changes in class distribution affect broad models. For better outcomes, adding economic time series data to these models will also be investigated.

For Potou zone saving and credit mutual in Senegal, (Kinda & Achonu, 2012) also attempted to provide a helpful scoring model to determine credit risk. The article's introduction stressed the importance of granting access to loans to the development of Africa. Despite the growing threat posed by the region's lack of loan availability, there is still a challenge in managing credit risk. It also brought up Senegal's rising bad debts as a result of poor credit risk management. The Potou zone in Senegal was the only area covered by the study for its intended purposes.

According to the definition provided in the article, credit scoring is the process of determining how well future loans will perform by using information about the characteristics and performance of previous loans (Schreiner, 2003). The article further stated that applying a good scoring model would help predict a person's ability to repay a loan. It explained that there are three scoring models: the statistical tree, the expert system, and the principle of regression. The statistical tree offers a borrower's risk level based on past experiences and characteristics. The expert system depends on the judgement of the officer in charge of loans. He explains that the loan officer uses their knowledge and experience to predict a borrower's credit worthiness. The principle of regression, as defined in the article, is based on sound mathematical formulas to establish a client's credit worthiness.

The article explained that the regression principle is the best scoring model due to its efficiency, predictive power, and operational capacity. Hence, the principle of regression would be used in the case of MECZOP. To further support the case for the principle of regression, it outlined three benefits. They asserted that using a regression model in assessing a person's credit worthiness helps the lenders in their decision-making process, the administrative process and at the cost level by reducing transaction costs. They explained that the model reduces human errors and provides a close to accurate predictability of a person's credit worthiness. It also helps the administrative process by providing an efficient analysis and increases the loan officer's sense of security in the organisation. They further explained that, although the regression model has its benefits, it also has some disadvantages. It must be noted that the goal of the model is to help decide between good and bad debt.

In developing the model, they grouped the data into dependent variables and independent variables. The article chose late repayment as its dependent variable. It further grouped its independent variables into socio-economic characteristics of the borrower, characteristics of the loan, and the loan officer's experience. A sample size of thirty borrowers was chosen within the period from 1st January 2007 to the 31st of December 2010. The data was then regressed and a regression model was created after further analysis. The model recorded a sensitivity of 86.67% and an overall accuracy to predict a delay in payment was over 90%, with a threshold of 0.5.

They explained that MECZOP's operations are funded by lending, hence, MECZOP should position itself to avoid bad debts. The model recommended that MECZOP keep an eye on borrowers below the age of forty-seven. This is because such people, according to the model, have a high probability of late repayment of loans. Additionally, they suggested that since women are more likely to repay a loan, the portfolio allocation of loans made to them should be raised. Additionally, MECZOP has to focus on loans that may be paid back in instalments. Collateral lowers the likelihood of non-payment; it is crucial that the MECZOP, in addition to the compulsory savings, requires collateral for high loans. The MECZOP would gain by being hesitant to extend credit to customers who have a history of late loan payments.

The majority of academics were likewise in favor of the analysis of credit risks using well-organized credit scoring systems. Among them are the studies “Credit Risk Measurement: Developed over the Last 20 Years” by (Altman & Saunders, 1996) which sought to track significant changes in credit risk measurement over the previous 20 years using credit scoring in the context of a discriminant analysis model and show how many of these changes were reflected in published papers.

Five possibilities are clearly outlined in the paper’s initial section on credit risk management. In order to determine the credit risk associated with corporate loans, the majority of financial institutions reportedly relied almost entirely on the first method, expert systems and subjective analysis, for the past 20 years. In order to make a mostly arbitrary decision regarding whether or not to extend credit, bankers used the so-called 4 “C’s” of credit. In the second system, known as an accounting-based credit scoring system, a financial institution’s (FI) decision-maker assesses potential borrowers’ various important accounting ratios against industry or group standards. To create either a credit-risk score or a probability of default metric, important accounting data are integrated and weighted. The study notes that there are at least four methodologies for creating credit scoring systems, including the linear probability model, the logit model, the probit model, and the discriminant analysis model, which has been by far the most prevalent approach in terms of JBF publications.

The third choice in the paper's section on credit risk management takes other (more recent) credit risk measurement methods into account. The “risk and ruin” models, a class of bankruptcy with a strong theoretical underpinning that, in many ways, resembles the option pricing models (OPM) of Black Scholes, Merton, and others, are the first of these models to be investigated (Hull and White, 1995). The paper goes on to say that the factor analysis method outlined in the study should be used to estimate credit risk in commercial banks. In addition, the capital market-based model, which is the mortality rate model of Altman (1988, 1989), and the ageing technique of (Asquithet al., 1989) are closely examined and discussed. A fourth, even more recent method is examined: the use of neural network analysis to solve the credit risk classification problem. These more recent models are all thoroughly discussed.

The authors claim that the expansion of off-balance-sheet instruments over the past 20 years has been one of the most significant developments. The research also examines the metrics of credit risk associated with these instruments. In the article, metrics of credit concentration risk are covered in great detail. Since the groundbreaking work of (Markowitz, 1959), Fixed Income Portfolio Analysis has allegedly made it possible to apply portfolio theory to ordinary equities. Concepts like individual stock and portfolio betas can be used in this study to highlight risk levels. Since most bank and bond portfolio managers have been unable to use this technique, the paper continues by outlining a process that will steer clear of significant data and analytical pitfalls that have dogged fixed income portfolio efforts and provide a sound and empirically workable portfolio approach.

The research also examines the return-risk concept. The standard mean variance of return framework is not suitable for long-term, fixed income portfolio strategies, the authors assert about the return-risk framework. However, when utilizing the equation where EAR stands for expected annual return, YTM for yield-to-maturity (or yield-to-worst), and EAL for estimated annual loss, measuring expected portfolio return for fixed income bonds and loans is actually rather simple. Additionally, the article employs this equation to evaluate bonds depending on their initial (or current) bond rating. The discussion of the study's findings in the report is excellent. The authors employed proxy risk indicators since they acknowledged that it was more challenging to evaluate projected returns for commercial loans. The equation for the conventional mean return-variance portfolio framework, which is used to calculate the conventional efficient frontier, is also supplied. The article also looked at the idea of using returns or durations to calculate return correlations due to the lack of historical high yield bond return and loan return data. The study concludes that using the variance of returns as a measure of the risk of either the individual assets or the portfolio is just unsuitable (both conceptually and practically). It is suggested that "unexpected losses" be utilized as the major indicator when establishing the requisite reserves against bank capital because many banks employ the RAROC (risk adjusted return on capital) approach. : As a result, it is recommended that each loan or bond that could be added to the portfolio be assigned a bond rating equivalent using a variation of the z-score model called the z-score model (Altman, 1993). The results of using this method are thoroughly discussed in this study".

The study provides and explains a method for measuring portfolio risk. This formula is used with data to get findings detailed in the paper and their analysis. On the identical bond portfolio examined throughout this work, a portfolio optimizer programme is run, but expected and unexpected losses rather than returns are used this time. Although they made it apparent that these are preliminary results, the authors claim that the results of the portfolio optimizer programme are consoling in that the unexpected loss produced by the z-score provides an alternate risk assessment.

Overall, the paper succeeded in achieving two goals. First, it demonstrated how credit risk measuring approaches had evolved over the previous 20 years and how many of these advancements were reflected in articles that had been published in the Journal of Banking and Finance (JBF). Second, a novel method for calculating the return-risk trade-off in a portfolio of risky debt instruments, including bonds and loans, was created in the study. The paper ends with a confident prediction that over the next 20 years, significant advances in data bases on historical default rates and loan returns will lead to the development of fresh, innovative methods for quantifying the persistent credit risk issues that financial institution (FI) managers face.

(Bachmair, 2016) also aimed to propose a four-step credit scoring structure for analyzing and measuring credit risk along the following factors: analyzing risk drivers, defining key characteristics to determine the choice of risk analysis approach, quantifying risk, and applying credit risk analysis and quantification to the design of risk management tools. The four-step suggested framework is based on analyses of academic approaches and has applications in the real world. It is applied in countries like Colombia, Sweden, and

Turkey. As demonstrated and discussed in the article, the proposed framework was put into practice in Indonesia, where it helped establish a framework for managing credit risk for sovereign guarantees.

The government is exposed to certain significant risks as a result of the contingent liabilities resulting from guarantees and the contingent assets resulting from on-lending, as is evident from the abstract. In order to accomplish specific policy objectives, the paper claims that government loans and sovereign credit guarantees can encourage private sector investment. But cautious risk management techniques and research can help in locating and lessening both current and potential dangers. The techniques used to manage credit risk must be extremely exact and founded in data from academic research, business practices, and international experiences. The peculiarities of the guarantee and on-lending portfolio, the specific risk exposure of the sovereign, the accessibility to market information and data, and the availability of resources and capabilities in the public sector were noted as important differentiators. According to the paper, the government should consider these strategies as iterative and long-term because creating a good framework for risk analysis and measurement necessitates a significant investment of time, money, and resources.

The literature reviews from academic and intellectual works on credit risk analysis were given weight in this piece. A well-defined risk management objective, a risk analysis, and the creation and implementation of a risk management strategy that includes monitoring, reporting, and reassessment procedures are all part of a strong risk management framework, according to (Anderson & Abousleiman, 2011). According to (Lewis & Mody, 1997), the government of Columbia entered into a Public and Private Partnership agreement in the early 1990s to encourage private sector involvement in important infrastructure projects like electricity generation, toll roads, and telecommunications. The government guaranteed specified project risks, especially demand risk, to entice investors. A recession in the economy in the late 1990s led to the activation of many guarantees. By 2004, total government expenditures amounted to 2% of GDP (Cebotari, 2008).

In this work, enough quantitative and qualitative data were used and evaluated; the researcher paid close attention to the data quality. According to the researcher, credit scoring requires more qualitative data, while simulation modules need a lot more quantitative data. Regarding risk variables and incidents, the researcher used historical data.

The Basel Committee has provided financial institutions with a framework for internal ratings-based (IRB) risk analysis (Basel Committee on Banking Supervision, 2005). The researcher suggested using a credit score approach to analyse the credit risks connected to the data obtained.

Financial ratios, competitive environments, and the regulatory environment were a few of the observable elements that the researcher looked at to determine credit quality. By regressing historical defaults on a few historical parameters, the statistical model was estimated. The performance of the chosen model was evaluated using the default probability, or the probability that the default will occur.

The study found that sovereign credit guarantees, and on-lending can significantly alter the balance sheets of governments in terms of contingent liabilities and assets. An integrated framework for risk management should be used to manage the risks associated with these contingent liabilities and assets. This enables the government to share the risk of significant investments made in the private sector while maintaining reasonable risk levels. One of the main pillars of such a risk management framework is the analysis and measurement of credit risk from guarantees and on-lending. Additionally, comparing different policy alternatives for government support is made easier by levelling the playing field by making costs associated with guarantees and on-lending public and transparent.

Similar to this, the foundation of banking is the concept of making money by lending money to people who need it. The banks will then take interest off of the borrower's payments. The likelihood of certain borrowers defaulting on their loans, which causes a loss to the bank, resulting in that loss. Therefore, banks determine a potential borrower's creditworthiness based on their likelihood of defaulting. As a result, determining credit risk is a fairly active area of research.

The goal of the thesis is to examine whether logistic regression can outperform the current heuristic credit rating model used by the cooperating corporate bank. This thesis' major objective is to clarify to the reader how credit rating models are made.

Starting with the definition of credit ideas, the thesis gave an outline of several notions. Afterwards, a general introduction of credit modelling that mentioned the Basel II agreement was given. Additionally, a thorough assessment of the literature on the Subprime Mortgage Crisis and the creation of credit rating models was conducted. Numerous references were used throughout the book as the researcher gave a thorough overview of the literature on certain widely used credit assessment models.

The researcher (Einarsson, 2008), employed data from the current corporate bank credit rating model, known as Rating Model Corporate, in the modelling procedure (RMC). Additionally, the data that is now available can be divided into qualitative, quantitative, customer, and other elements and statistics. Data analysis, model construction, and result validation use statistical or mathematical techniques. We introduce and apply linear models and generalised linear models. Similarly, a method for reducing multidimensional data sets to lesser dimensions, discriminant analysis, various classification techniques, fundamental statistics principles, and more sophisticated statistics approaches were also introduced.

Additionally, validation techniques were employed to gauge how well each modelling approach performed. As well as the crucial idea of selective power, ROC curves and visual representations of relative and cumulative frequencies are introduced.

Conclusions of the thesis revealed that of all the modelling strategies investigated, logistic regression models were chosen to be the most realistic way for simulating default probability. Additionally, it has been found that the linear and quadratic discriminant analysis methodologies are not applicable for the modelling of credit default since both of them need normality of the predictive variables, which makes it challenging to include customer features. The researcher also finds that the amount of data available cannot be regarded as optimal, not only due to the relatively low number of defaults and the fact that only three years can be considered during the modelling process, but also because the variable selection analysis was severely lacking due to the absence of various quantitative key ratios. Finally, it was found that there does not seem to be a single numerical indicator of the credit rating model's efficacy in the validation operations.

According to the theory, it would be extremely intriguing to build a neural network credit rating model, one of which would involve taking into account different macroeconomic data. As Altman and Saunders note, the thesis also concludes that it would be beneficial to include fixed income portfolio analysis in the analysis. After reading the study, it ought to have been better organized and grammar-checked for errors. Wikipedia is also not a trustworthy source for a research paper. However, this research project was extremely thorough.

More specifically, the goal of the study by Khandani et al. (2010) on "consumer credit-risk models via machine-learning algorithms" was to create a model that could foresee consumer credit events using machine learning approach months in advance. The introductory paragraph of the article emphasized how important consumer spending is to macroeconomic conditions. The US had approximately $13.6 trillion in outstanding consumer debt in 2008, though. Consumer credit was obviously a problem that required attention. Instead of using human judgment, it is necessary to rely on models and algorithms to create a scoring model that forecasts a consumer credit event. The report also mentioned that it was typical for lending organizations to create a model using the personal information of borrowers. It is made clear, though, that these models only offer a modest degree of accuracy and evolve slowly. A fundamental metric for evaluating the risk of consumer credit was proposed in the paper, which considers both a customer's banking activity and debt to income ratio. It asserted that by examining a customer's bank transactions and debt payments, it can more accurately determine a consumer's credit worthiness than other models. The article's methodology was the machine learning approach.

From January 2008 to April 2009, transactions and credit bureau data from a sizable commercial bank will be combined in this strategy. The dataset used in this article includes information on individual customers' account balances, credit bureau reports, and bank transaction levels. Between January 2005 and April 2009, data was collected from clients of a commercial bank. Due to privacy protection laws, names, addresses, and social security numbers were deleted. The transaction data were broken down into channels and categories to help with the data analysis. We discovered resources like ATMs, online bill-paying, etc. as well as categories like eating out and drinking out, among others. In addition to account transactions, the bank added credit file information provided by one of the credit bureaux for its customers. Also included in the data were each customer's available credit or loan options across all financial institutions. The banks reidentified every consumer. The calculations were carried out using the bank. Credit card banking and payment habits were the main subjects of the article. The top 5% of the sample was discovered to have a credit card debt that was approximately $20,190 at the beginning of the sample and increased to $27,612 by the end.

In the next paragraphs, the article described the machine learning technique used to create a forecast model for the bank's clients and transactions from January 2005 to April 2009. In the modelling, concepts like income shocks were included. Income shocks are calculated as the difference between the current month's income and the six-month moving average of the past six months' income, divided by the average period's income's standard deviation. We create a credit risk forecast model that forecasts defaults for specific customers using the machine learning methodology. The model discusses two crucial topics: how it differs from previous models and how prevalent credit and delinquencies are in the economy. The forecast model in the paper is built using regression trees. The machine forecast used a fair baseline based on credit score. However, it is noteworthy that the machine learning model predicts significant predicted delinquencies and default risk for clients with higher credit quality. This is so because machine learning forecasts contain information that is distinct from that in conventional credit score models. According to the publication, not all consumer spending and saving activities are included in the dataset.

The data set was divided into equal-sized samples and then segregated by the availability of features in order to test the model's resistance to missing features. They then carry out a cross validation by 10 times. It has been found that a few features make it simpler to predict account delinquencies. The article proposes a novel way to quantify system risk by combining individual accounts to generate a forecast of consumer credit delinquencies. Despite being a small subset of the bank's customers, the data enables the machine learning model to accurately capture the population's evolving risk characteristics.

In conclusion, a model that can forecast credit even 3–12 months in advance is necessary. This is made easier with the use of machine learning. The outcomes are seen to sign a more potent model of consumer behaviour.

The aim of (Jackson & Perraudin, 2000)'s study, "Regulatory implications of credit risk modeling," was to draw attention to the numerous shortcomings of credit risk assessment models and offer recommendations for how to improve model evaluation through the use of back testing and other advanced techniques.

Both portfolio-theoretic mark-to-market and default-mode models can be used to estimate credit risk, according to the researchers. The aforementioned models were created to assess the risk involved in holding credit-sensitive securities. By forecasting the distribution of portfolio value at a given time and allowing for a decline in credit quality even in the absence of a total default, mark-to-market models generate measures of portfolio value. Over a given horizon, the "Value-at-Risk" (VaR) and default-mode models are used to estimate the distribution of total defaults on exposures in the portfolio". In accordance with their findings, the researchers also recommended categorizing credit risk assessment algorithms.

The researchers concluded that changes in the debt and loanable funds markets were responsible for the emergence of the credit derivatives market and the unprecedently rapid expansion of the already-existing loan and securitization markets. Banks and other financial institutions eager to accept and put credit risk-mitigation measures into practice found the new credit risk assessment models to be of great value. The new models can be used to calculate regulatory capital and correctly priced the exposure portfolios included in securitization. According to the Basel Accord's regulatory framework, held-up capital had to be at least 8% of a company's exposure to the private sector. Three factors have made these developments cause concern. Firstly, as banks reduce their low-risk exposures, the average riskiness of what is left will increase, lowering the effective capital required under the Basel 8% limit. Secondly, Regulatory arbitrage operations frequently lower capital requirements without significantly lowering the bank's risks. Finally, complex regulatory arbitrage transactions make bank activities less transparent to market participants and regulators.

The researchers initially took parameterization by judgement, risk omission, data blanks, and model performance into account when addressing the flaws in credit risk models.

The parameters that affect the calculated risk measures but may be considered by the user while using the model are among the problems attributed to the models. According to the researchers, parameterizing the model based on historical data on current dividend pay-out rates may produce inaccurate results.

The current risk assessment models are identified with some omissions or consider the relevant correlation between different risk categories as zero. Interest rates and other market risks are left out due to simplifying credit risk models by current models. Blank data was also recorded as a problem faced with current risk models.

The effectiveness of some models for assessing credit risk was examined, and some of the studies gave simulation results on a variety of models, focusing primarily on the degree to which different models produce comparable risk outcomes for the same portfolio at the same time.

Internal credit risk models are essential for banking monitoring, according to the researcher. In the study, back-testing credit risk models was also considered, and it was discovered that this method was more difficult than back-testing market risk VaR models. To manage the issues that come with models, it is important to distinguish between (1) standard models that are applied to publicly available data, (2) standard models that incorporate bank-specific data, such as internal ratings, and (3) internally built models that use in-house data. Through cross-bank comparison, heterogeneous bank models were put to the test. Benchmarking, according to the experts, would allow for more rigorous comparisons across different banks.

The report argued that because of their shortcomings, credit risk models should not currently be the only element considered when determining regulatory capital for credit risk. Supervisors, however, should be aware of the less demanding roles that credit risk models may play in the monitoring regime. Supervisors, like senior managers, can contribute to the process of deciding on the standard of risk management and the appropriate capital levels for a particular bank by learning a lot from the outcomes of credit risk models.

If supervisors are to have confidence that banks are utilizing their model wisely, designing the framework for back-testing, cross-bank comparisons, and penalty mechanisms will be a critical step in the process. In their article “Evaluating Credit Risk Models,” (Lopez and Saidenberg, 1999), recommend using a panel data strategy as a way to assess cross-sectional simulation-based credit risk models. In order to more accurately identify the financial risks, they confront and allocate the necessary financial capital, banks are said to have expended a lot of resources over the past ten years in developing internal risk models. The claim is supported by the Basel Capital Accord’s 1997 Market Risk Amendment (MRA). There are further mentions of Credit Suisse Financial Products (1997) and J.P. Morgan (1998) as examples of the credit risk modeling field’s quick development. The study uses the “International Swap Dealers Association (ISDA, 1998) and the Institute of International Finance’s Working Group on Capital Adequacy (IIF, 1998) to show the different concerns with credit risk models.

The paper cites two organizations as proof that there are numerous “credit risk models that vary in their necessary assumptions, such as how credit losses are defined, including the Basel Committee on Banking Supervision (BCBS, 1999) and the Federal Reserve System Task Force on Internal Credit Risk Models (FRSTF, 1998). The Federal Reserve System Task Force on Internal Credit Risk Models (FRSTF, 1998) and the Basel Committee on Banking Supervision (BCBS, 1999) are still referenced under the general issues in credit risk modelling, and it is acknowledged that there are two sets of critical issues that must be resolved before credit risk models can be used to determine risk-based capital requirements”. The inputs to these models are the subject of the first set of issues, whilst the specification and validation of the models are the subject of the second set.

The essay goes into greater detail on the subject of evaluation techniques utilizing simulated credit portfolios. Here, “stress testing,” a technique for evaluating credit risk models, is covered. Using this method, the effectiveness of a credit risk model is evaluated in respect to event scenarios, whether they are based on hypothetical results or real-world outcomes. Researchers have begun comparing the forecasts from various credit risk models, and references are made to (Crouhy & Mark, 2000; Gordy, 2000; Koyluoglu & Hickman, 1998) as examples. The section titled "Intuition from time-series analysis" of the paper elaborates on how methods typically used for forecast evaluation and model specification, whose general goal is to ascertain whether a series of out-of-sample exhibit properties typical of accurate forecasts, can be modified for use with panel-data analysis, such as credit risk modelling (Granger and Huang, 1997).

The researchers also recommended employing simulation methods to obtain the additional data of credit portfolio losses required for model assessment. They can use this to their advantage by analyzing credit risk models. In order to create additional credit portfolios, the research uses a novel methodology that only requires resampling and replacement from the initial panel dataset of credits. The strategy used in is similar to (Carey, 1998). The article first creates resampled credit portfolios in order to produce the anticipated cumulative density functions of credit losses. The essay focuses on three alternative hypothesis testing methods. Using the idea of Mincer-Zarnowitz regressions, the initial method for testing a hypothesis involved comparing the expected losses predicted by the model to the actual losses recorded on the resampled portfolios. The study recommended utilizing the likelihood ratio statistic and the sample's overall number of exceptions to assess whether the proportion of observed exceptions is equal to 1 in order to test the second hypothesis. The last hypothesis testing strategy of the study was applied to examine whether observed quantiles obtained from a model's distribution forecasts display the properties of observed quantiles from correct distribution forecasts.

The paper also cited (ISDA, 1998) to emphasize that there are different levels of sophistication for credit risk models. The study goes on to discuss the disadvantages of using the suggested: simulation approach, which permits the comparison of a model's predicted credit loss distributions to observed results, as in the typical back tests performed for VaR predictions. The first restriction stated that changes in the weights of a credit portfolio, changes in credit quality, and changes in the value of credits of a certain quality are the three main factors that determine changes in a credit portfolio's worth (possibly due to changes in credit spreads):. The second problem was that credit risk models—which were mostly panel data models—were only useful for datasets in which the ratio of assets to years was significantly higher. The third mentioned restriction is that the resampling method is restricted to the initial set of credits.

The paper concludes that due to their different underlying time periods, credit risk model evaluation will always be more challenging than market risk model evaluation. The assessment methods suggested give performance evaluation in a cross-sectional situation and are based on statistical resampling, as indicated in the research's overall conclusion. Additionally, these suggested tools are quite easy to use. According to the report, some features of the recommended evaluation method need more study, and future work should focus on direct comparisons of credit risk models across different credit datasets. Furthermore, "Credit Risk Management of Ghanaian Listed Banks" by (Apanga et al., 2016) uses well-structured accounting credit score content analysis techniques to investigate the credit risk management practices of Ghanaian financial institutions and compare them to Basel II (1999) requirements. The article claims that financial organizations are subject to a range of hazards, including market risk, credit risk, and operational risk, suggesting that utilizing risk management strategies is unavoidable. Banks and other businesses in the financial services industry consider the control of credit risk to be one of these risk management techniques to be particularly important. The authors' assertion that credit risk management is the most important risk management strategy is regrettably unsupported by any research or data, while appearing to be true. Nevertheless, it is asserted by (Fatemi & Fooladi, 2006) who support the notion that credit risk management plays a crucial role in the overall risk management operations carried out by organizations in the financial services industry.

A comparison-based component is also present at the article's start. By referencing examples from the corporate finance literature, the authors let us know that credit risk research in industrialized nations is well established (Altman & Saunders, 1998; Fatemi & Fooladi, 2006). However, despite the fact that the debate of credit risk in policy talks is becoming more important, research is still lacking, especially in developing countries. In instance, there hasn't been much research on credit risk management difficulties in Sub-Saharan Africa, and Ghana in particular, despite the notion that the current financial woes in the banking sector pose a severe threat to the health of both developed and emerging economies. To fill this vacuum, the authors of this research look at the credit risk management practices of Ghana's listed banks.

Additionally, the paper compares Basel II to the credit risk management practices of Ghana's listed banks (as at end of 1999).

A researcher is already aware that sufficient and pertinent literature is needed in order to do this type of research fully and properly. This essay provides an overview of the banking industry in Ghana by thoughtfully and attractively integrating important information. According to (Aryeetey & Gockel, 1991), Ghana's current financial system was created as a result of the Economic Recovery Programme, the Financial Sector Adjustment Programme (FINSAP) that followed, and the Financial Sector Strategy Plan (FINSSP), all of which were implemented in the early 1980s to revive the country's ailing economy. The Financial Sector Adjustment Programme (FINSAP) and Financial Industry Strategy Plan (FINSSP), according to (Quartey, 2005), were created with the intention of liberalizing the financial sector in order to improve bank efficiency and boost public confidence in the sector.

There are 17 principles for managing credit risk, according to a review of Basel II literature, which increases the burden on businesses to implement effective credit risk management practices. The Basel Committee on Banking Supervision (1999) (BCBS), which was founded by the Bank for International Settlements, is clearly and rightfully credited with developing these principles (BIS). The division of the 17 credit risk management principles into the three main thematic categories of organizational structure, operations/systems, and policy and strategy is also reinforced by (Kannan, 2004). The authors' analysis of the literature leads them to the sound conclusion that banks need effective credit risk policies and procedures to direct their loan-granting and credit risk management operations. In order to effectively establish and implement these policies and objectives, they also require well-structured organizational structures that define the specific functions of directors, executive management, and other staff members. In order to reduce the frequency of loan losses, banks must also maintain systems that enhance efficient credit administration, credit risk measurement, and monitoring operations.

Part of the case study technique used for the study in this article is doing an empirical analysis of a specific contemporary occurrence in its actual context. So, Case 1, Case 2, Case 3, and Case 4 are the names of the banks employed in the study. Case 2 is the smallest bank among all the listed banks in Ghana as of 2007, Case 3 is a traditional foreign bank with a presence around the world, and Case 4 is the only bank in this study whose branch count has not increased. Case 1 is the largest universal bank in the country in terms of branch networks. Surveys, manuals on internal credit policies and procedures, semi-structured interviews with credit risk managers of the selected banks in May 2007 and October 2014, and conversations with them are only a few of the many sources used to obtain information. Based on the theme areas of the study, the interview guide is designed to gather data from the key participants (credit risk managers and officials) of the four listed banks. The credit risk manager is interviewed in May 2007 at the corporate offices of each bank. More importantly, interview participants are given the option to remain anonymous as a guarantee that the data acquired would be treated in the strictest secrecy.

The approaches employed analyzed responses in light of the study's thematic issues and provided a descriptive analysis of the findings for the examination of the findings. After analyzing the data, it became clear that the main sources of credit risk exposure were corporate and small business commercial loans, interbank transactions, trade finance, and currency exchange activities. The main types of credit risks that banks face, such as counterparty default risk, equity risk, securitization risk, concentration risk, and residual risk—although residual risk is only a modest problem for banks as a whole—are also covered in detail. Additionally, responses to interview questions demonstrated that collateralized loan obligations and guarantees are the most popular strategies employed by banks to reduce their credit risk exposures. Additionally, banks have developed credit risk measuring procedures (using experts and subjective analyses) that enable them to quantify the risks associated with exposures to particular borrowers or counterparties. The article extensively explains and supports these methodologies. In addition to character (reputation), capital (leverage), capacity (earnings volatility), collateral (security), and reputation, banks also use these metrics to grade their risk more subjectively.

The conclusion of the article is that, with the exception of the boards of directors' involvement in the development of credit risk policies and strategies, the actions of the banks are in accordance with the necessary criteria. The analysis of the interviewees' comments offers a comprehensive explanation for this, and it is amply demonstrated that senior management and the banks' risk management division prefer to create lending rules and procedures for the board to evaluate and approve. The primary cause of this is that each board is fundamentally lacking in the specialized knowledge required in fields like risk management. The article continues by stating that this distinction is not always a negative thing because banks are required to alter their operating processes to correspond with the extent, complexity, and nature of their lending activity.

The article also demonstrated that the banks' techniques for controlling credit risk adhere to industry standards. Those credit plans and policies developed and implemented by banks' well-organized risk management departments, which are staffed by highly qualified professionals who have received the boards of directors' approval and recommendations. Additionally, the banks have a manual of credit policies and procedures that they can refer to when conducting lending operations. However, where they diverge from one another is in how the boards of the banks define the permitted loan types and maximum maturities for the various sorts of loans the banks grant. The requirement that these regulations be written by individuals with in-depth expertise of risk management explains this distinction.

Although the authors of the study believe Basel II's adaption to be a bold step, they also note that it does not serve as a panacea for good credit management practices. They recommend that the Bank of Ghana actively support ongoing education of Basel II's implications and implementation among its banking industry participants, analysts, and the general public. If the loan recovery divisions of the banks' issue loan management systems and workout strategies comply with best practices, more investigation is needed. This is another another judgment that has been made. More precisely, the paper by (Kun & Duo, 2014), titled "Credit Risk Management of Commercial Bank," sought to identify and provide an appropriate solution to some credit risk management problems that Chinese commercial banks were currently dealing with at the time of this research. By examining both internal and external factors that affect the growth of credit risk, the study used credit scoring systems to analyze the degrees of credit risk associated with Chinese commercial banks.

The inability of debtors to continue making loan payments was defined by the researchers as credit risk. They also described credit risk in terms of the possible outcomes of bank crises brought on by the emergence of a sizable number of non-performing loans. According to researchers who spoke about the current state of credit risks associated with commercial banks in China, these institutions still deal with loan-related issues using antiquated ways, which has resulted in a number of very serious disadvantages on their side. The researchers recommended that Chinese commercial banks learn about and apply the responsible credit risk management practices used in the West. They recommended adopting Westerners' prudent credit risk management models by, for instance, carefully allocating the quality of commercial banks' assets, establishing, and strengthening the credit system, and so on. They gave examples of countries like the United States of America (USA) and others that have decades of experience dealing with credit risk-related issues.

The researchers found that the following factors contributed to the internal formation of credit risk among commercial banks in China: a severe lack of basic bank documents and credit archives, a flawed internal credit control system, a lack of mechanisms for preventing and warning about credit risk, and a lack of information between banks and businesses that could result in problems with non-performing loans. A few external factors were also noted by the researchers as potential causes of credit risk among Chinese banks. Inadequately designed and built credit systems, financial system delays that also restricted the prevention and control of possible credit risk, and excessive government intervention were all highlighted by the researchers as probable drivers of credit risk development. In the context of their risk management framework, risk management model, and the knowledge that other commercial banks may learn from Citibank's responsible credit standards, the researchers took into account the credit management guidelines of Citibank.

The study discovered that despite commercial banks' full knowledge of the significant risk associated with adhering to conventional credit management rules, they were nonetheless wary. Information asymmetry, which is frequently believed to contribute to the development of credit risk, did not provide a challenge for Chinese commercial banks. Following extensive data collection and analysis of the internal and external factors that contribute to the development of credit risk among Chinese commercial banks, the researchers presented the following warnings and recommendations: staff members' work capacities should be improved, and adequate staff quality enhancement was highlighted; commercial banks were advised to enhance their credit risk warning mechanisms and to establish and improve a special ad hoc committee. Additionally, the study found that actions like mortgage inspection, scientific and reasonable loan risk assessment, liquidity planning, and rigorous adherence to responsible credit risk management guidelines could all contribute to enhancing commercial banks' credit ratings. In addition, the relevant regulatory bodies should improve credit laws and regulations, strengthen oversight of commercial bank loans, increase borrowers' risk awareness, tighten political controls, and other related things.

The goal of Kun and Duo’s “Credit Risk Management of Commercial Banks” essay from 2014 is to educate readers on the origins and progression of credit risk management in Chinese commercial banks. The essay is divided into four sections: commercial banks’ credit risk, Chinese commercial banks’ current credit risk environment, problems with China’s credit risk management, and finally, measures for mitigating rising credit risk.

Principles of Credit Risk Management

The evaluation of credit risk analysis must pay enough attention to the “Principles for the management of credit risk” of the Basel Committee, which seek to offer detailed guidance on how credit risk can be managed successfully in our current global economy. The introduction and the guiding principles for evaluating banks’ credit risk management were therefore the two main components of the essay. The July 1999 initial release of this document for discussion had an impact on the development of this current edition. The essay underlines the significance of effective credit management and points out credit risk as one of the main issues facing banks throughout the world.

It was made apparent in the Introduction that there was a need to improve global credit risk management. “The credit status of a bank's counterparties may worsen in some cases due to a variety of factors, including lax credit criteria for borrowers and counterparties, ineffective portfolio risk management, failure to pay attention to changes in economic conditions, and others. The danger that a bank borrower or counterparty will not fulfil its obligations in line with the terms stipulated in the agreement is described as credit risk in the article. Despite the fact that there are other sources of credit risk, such as acceptances, interbank transactions, trade financing, foreign exchange transactions, financial futures, options, bonds, equities, and bonds, as well as the extending of commitments and guarantees and the settlement of transactions, loans are said to be the biggest and most obvious source of credit risk. The essay emphasizes once more how crucial it is for banks to learn from their past mistakes and meticulously track and manage these lending-related risks. The article's better procedures, in the committee's opinion, will have an effect in four crucial areasThey were cultivating an environment that was advantageous for credit risk by carrying out a responsible credit-granting process, implementing a suitable credit administration, measuring, and monitoring approach, and making sure there were adequate controls over credit risk. The essay makes it apparent that an extensive credit risk management program will solve these four difficulties, even though other countries and even institutions may employ different strategies. It should be underlined that the actions of the bank should be taken into consideration when implementing a credit risk management approach.

The board of directors was requested to periodically (annually) keep reviewing the bank's fundamental credit risk policies and credit risk strategy in order to provide an optimal credit risk environment. The approach should be in line with the bank's propensity to assume risk and the amount of revenue it anticipates from assuming various credit risks. The board of directors' approved credit risk strategy must be implemented, and senior management is in charge of creating the procedures and policies for identifying, evaluating, managing, and controlling credit risk. Banks are encouraged to utilize extensive, distinctive credit-granting criteria in order to operate under a favorable credit-granting process. These standards should clearly identify the bank's target market in addition to supplying details on the borrower or counterparty, the credit's objective and structure, and the source of the loan's repayment. To provide a transparent credit giving process, banks should generate general credit limitations at the level of particular borrowers and counterparties, as well as groups of connected counterparties. These restrictions should aggregate various exposure types in a comparable and useful manner, both in the banking and trading book as well as on and off the balance sheet. For granting new loans as well as revising, refinancing, and renewing existing credits, banks should have a clear procedure in place. Credit must always be given in an impartial manner. Loans to related firms and individuals in particular must be approved with special consideration, closely supervised, and additional safeguards must be taken in order to limit or lessen the risks associated with lending agreements that are not at arm's length.

Banks are urged to have a competent system for the continuing management of their various credit risk-bearing portfolios if they want to maintain a strong credit administration, measurement, and monitoring procedure. Banks are also asked to put up a mechanism for monitoring the status of specific credits. To manage credit risk, banks are urged to create and use an internal risk assessment system. The rating process must be consistent with how the bank operates. All banks must have a system in place for gaining access to the overall framework and standards of the credit portfolio. Furthermore, it is suggested that banks evaluate their credit risk exposures in challenging circumstances and consider upcoming economic changes when evaluating individual credits and their credit portfolios.

Banks must implement a system of objective, continuing review of the bank's credit risk management policies in order to properly manage credit risk. This means that in order for the board of directors and senior management to come up with fresh suggestions for reducing credit risk, a constant evaluation of the bank's credit policy should be examined and conveyed to them. Banks must also make sure that their operations for extending credit are properly managed and that credit exposures are kept within bounds that are consistent with reasonable standards and internal constraints. In order to make sure that any deviations from policies, procedures, and restrictions are reported promptly to the prudent level of management for action, banks should develop and implement internal management and other practises. Banks need to set up a system for dealing with issue credits, early loan remediation, and other workout circumstances.

Supervisors should make sure that banks have a productive system in place to detect, analyze, monitor, and control credit risk as part of a more comprehensive risk management strategy. Supervisors should objectively evaluate a bank's loan-giving and ongoing portfolio management policies, rules, practices, and strategies. The prudential constraints that regulators could impose on banks' exposure to specific borrowers or networks of connected counterparties are worth considering.

The article's goal by (Raad, 2015) was to support the impact of sound credit risk management procedures on the firm's profitability and sustainability. The article established a link between the bank's performance and credit risk management practices by thoroughly analysing Basic Bank Ltd.'s (BBL) credit risk management methods and policies. The study started by proving that risk is an inherent component of banking. In addition, he claimed that although there are many business hazards, credit risk makes up around 50% of all business risks worldwide. Consequently, the organisation's success depends on establishing sound credit risk management procedures.

The BBL's credit risk management procedures were also covered in great depth. In order to help them, it was reported that BBL has put different credit principles into place. A couple of these rules included the following: no term loan would be permitted for the commercial sector, and total loans should not be more than ten times the bank's net worth. The bank divided its tasks into corporate, SME, retail, and credit card obligations to be carried out by its employees. The credit approval team, asset operating department, recovery unit, and impaired asset management were the four teams that participated in the lending process. The diverse policies and philosophies of the banks served as the teams' guides. On the banks' loan recovery policies, it went on in great detail.

It should be highlighted that BBL's loan and advance volume surged throughout the course of the past five years (2008–2012), with 2012 seeing the biggest number of loans given as a result of poor management. Except for 2012, unclassified loans at BBL climbed at an even rate during this time, according to the report. According to the ratio research, the ratio of standard loans to total loans climbed in the first four years but declined in 2012. Over the first four years, the classified loans to total loans ratio fell once more, although it slightly rose in 2012.

The purpose of the paper was to determine how credit risk management techniques affected the bank's financial performance. The annual

BBL report provided the sample data (2008-2012). The non-performing loan ratio, capital adequacy, and credit loss provision were chosen as the paper's independent variables, while the return on asset was regarded as its dependent variable. In order to conclude, these variables were subjected to numerous regression analyses.

Over the previous five years (2008–2012), BBL has seen a growth in loans and advances, with 2012 having the largest number of loans provided due to subpar management techniques. According to the paper, with the exception of 2012, unclassified loans at BBL grew steadily over this period. Over the first four years of the ratio analysis, the standard to total loans ratio rose; however, in 2012, it fell. The percentage of classified loans to total loans declined once again throughout the first four years, but it slightly increased in 2012.

The study aimed to ascertain how the bank's profitability was impacted by its credit risk management practices. The representative data was derived from the BBL annual report (2008-2012). The return on asset was chosen as the paper's dependent variable, with the non-performing loans ratio, capital adequacy, and credit loss provision chosen as its independent factors. These variables underwent various regression analyses in order to get a result.

Credit risk management practices were discovered to be essential to the success and long-term viability of bank operations. Additionally, it was established that the biggest barrier to credit risk management is the expense of employee incentive. Further investigation revealed that the BBL follows the Bangladesh Bank's requirements when it comes to credit risk management. It frequently focuses on industrial credit policy rather than general credit policy.

The asset management, operation, approval, and recovery units make up the BBL CRM department.

It advised BBL to establish a detailed written policy for credit risk management procedures. Additionally, it must include a credit rating system to spot borrowers who might be high-risk. Once more, it was advised that BBL get adequate collateral before making loans or advances.

In order to reduce defaulted loans, the report concluded that it is critical for BBL to be able to recognise possible risk borrowers.

Due to Basel Committee on Banking Supervision and Central Bank of Kenya rules, commercial banks in Kenya must employ a variety of risk management strategies in order to satisfy their performance standards. Studies show conflicting results on how risk management affects commercial banks' performance. The studies conducted, however, did not employ the same metrics of bank performance, such as ROI, ROA, and ROE, in addition to varied risk assessments, as shown by the literature review, which is why the results were inconsistent. Most of these investigations were conducted outside of Kenya in the distant past. As seen from the foregoing, there is still some debate on how risk management affects bank performance. Consequently, there was a need to close a gap and contribute empirically to the body of research.

By focusing on loss given default, capital adequacy, non-performing loans ratios, and loan loss provisions as the independent variables and bank stock performance as the dependent variable, the study by (Onang'o, 2017) aimed to close a gap and empirically add to the existing literature on the effect of risk management on the performance of commercial banks. More current data are also used in the study. A benefit of this study was the expansion of the empirical literature in this field due to the use of stock performance, which is slightly distinct from the conventional balance sheet-based performance metrics.

Finding out how credit risk management affects the financial performance of commercial banks listed on Kenya's Nairobi Securities Exchange was the study's main goal. The following goals were also established: (a) Describe the relationship between performance and the capital adequacy ratio of Kenya's commercial banks. Analyze how the loss given default ratio affects the performance of Kenyan commercial banks. (c) Look at how Kenyan commercial banks' operating results affect their ratio of loan loss provisions. Find out how Kenya's commercial banks operate in relation to the ratio of non-performing loans.

The researcher used a quantitative longitudinal research methodology and a positivist research ethic for this work. The capital adequacy ratio (CAR), non-performing loan ratio (NPLR), loan loss provision ratio (LLPR), and given loss default ratio were also employed as four independent variables in the study (GLDR). According to the performance of their stock prices, these risk management criteria have an impact on banks' financial performance (independent variable). A generalized least squares (GLS) model for panels with random effects was also employed to construct a regression equation. Ten banks were chosen as the study's sample size from among Kenya's four registered and operational banks. The information for these 10 banks was then gathered in the following stage from their annual reports for the seven years between 2008 and 2014. The majority of these data may be accessed digitally or on paper. A Hausman test was also performed to determine whether to employ a fixed effects model or a random effects model using the null hypothesis that random effects were preferred over fixed effects as the preferred model for the data. The Panel Cross-Section Dependence Test, Serial Correlation Test, Panel Cointegration Test, Multicollinearity Test, Heteroscedasticity Test, and Normality Test were additional tests that were performed.

The capital adequacy ratio (CAR), according to the data, was regularly distributed. The results of the Hausman Test also showed that it was appropriate to employ the random effects model for this data. All levels of significance indicated that the variables capital adequacy ratio (CAR), loss given default ratio (LGDR), loan loss provisions ratio (LLPR), and non-performing loans ratio (NPLR) failed the normalcy test. As a result, the power transformation approach was used to transform all of the original data into new, roughly normally distributed data." Additionally, it was determined by the Abnormal Returns Unit Root Test that the panel was stationary at level; by the Capital Adequacy Ratio Unit Root Test that the panel was stationary at level; by the Loss Given Default Ratio Unit Root Test that the panel was stationary at level; by the Loan Loss Provision Ratio Unit Root Test Results that the panel was stationary at level; and, finally, by the Non-Performing Loan Ratio Unit Test that the same result was found. Additionally, Multicollinearity Test results for each of the four independent variables showed no evidence of collinearity. The researcher came to the conclusion that the data did not exhibit heteroscedasticity as a result of the heteroscedasticity test. The dependent variable was lagged with a lag of one in order to remove the dependency after the cross-section dependence test revealed that the data had panel serial correlation. Additionally, a non-significant but favourable correlation between Kenyan bank stock performance and capital adequacy ratio (CAR) was discovered by the study. The study also discovered a slight but favourable correlation between Kenyan bank stock performance and Loss Given Default Ratio (LGDR). The study also discovered a weak and unimportant correlation between Kenyan bank stock performance and the Loan Loss Provisions Ratio (LLPR). Finally, the analysis discovered a substantial inverse association between Kenyan bank stock performance and the non-performing loans ratio (NPLR). However, the model as a whole was significant in predicting bank performance. The low adjusted R-squared value revealed that the model had little predictive ability in employing the independent factors to explain the dependent variable.

There was a succinct overview of the study. The overall goal and specific goals of this study were to determine the effects of credit risk management on the performance of commercial banks listed on the Nairobi Securities Exchange in Kenya by using variables such as capital adequacy ratio, non-performing loans ratio, loan loss provision ratio, and capital adequacy ratio.

Not all of the chapters had conclusions. A segment was also devoted to the study's general conclusion. The study's overall conclusion clearly and concisely stated all the important findings pertinent to the study's objectives.

In light of the study's findings, numerous recommendations were made. One was that, given the nonperforming loans ratio's large impact on bank profitability, credit risk managers should handle it with more caution.

Additionally, it was advised that the regulatory non-performing loan ratios be adjusted in order to lessen the negative effects while maintaining the current regulatory policy requirements for loss given default ratios, capital adequacy ratios, and loan loss provisions ratios because their results are consistent across the sample.

The study adequately addressed its goal and was basically well organised.

The unpublished doctoral study dissertation "Credit Risk Management Practices: A Comparative Study of State Bank of India and Punjab National Bank" by (Bhardwaj, 2013) claims that the Indian banking sector has suffered significant losses during the last 10 years. Due to credit exposures that changed to negative levels, derivative exposures, or interest rate positions taken that may or may not have been thought to be hedging balance sheet risk, businesses that had been operating normally abruptly declared substantial losses. Commercial banks have virtually universally started elevating their risk management and control systems in reaction to this. Investment choices are challenging. Companies typically lack the same level of diversification as investors, making survival a significant and valid goal. Financial and investment decisions should be made with a minimal probability of financial trouble in mind. This is due to the fact that financial hardship increases the cost of filing for bankruptcy, which arises from the structure of the bankruptcy procedure and practically continuously results in a drop in stakeholder value above and beyond the reduction that occurred as a result of unfavourable occurrences.

The goals of the study were understanding the various practises used by the State Bank of India and Punjab National Bank, researching the framework these banks have adopted, and figuring out the most crucial factors they take into account for their credit risk management practises and the significant effects of those practises. These goals were expressly stated, constrained by the researcher's abilities, and, most importantly, made plain by the study's title and problem.

The study's specific objectives are to (a) investigate the credit risk management frameworks of the State Bank of India and the Punjab National Bank, and (b) research and contrast the risk management practises of the two banks.

The researcher gave a summary of the many studies on credit risk management techniques that have been done. A thorough evaluation of the literature was supplied by the researcher, and the article had numerous citations. There was both theoretical and practical justification, and related studies were referenced. The researcher employed the correct APA style for reference citations, and the reference of citation appeared to be accurate."

The researcher employed a sample of sixty officers who were chosen at random from among respondents from different State Bank of India and Punjab National Bank branches. The geographical coverage, however, was limited to the Ludhiana branches exclusively. Based on respondents' willingness to share their responses and using a convenience sampling method, data was gathered. However, secondary data were gathered via the annual reports, journals, and websites, while primary data were gathered through a systematic, predesigned, and transparent questionnaire. On the gathered data, the following tests were applied: mean score, standard deviation, weighted mean, percentage, t-test, and chi-square test. The researcher then acknowledges the following limitations of his work and lists them: Due to time and resource limitations, only respondents from Ludhiana City were taken into consideration for the survey. Future research may use a sizable sample from several cities. Many respondents were not very familiar with the phrases used in the questionnaire; therefore, face-to-face interviews were conducted. According to the author, unavoidable biases may prevent respondents from providing information that is entirely correct and the connection between a government supplier's willingness to be knowledgeable about credit risk management techniques could not be thoroughly investigated.

The results showed that, according to the information gathered by the State Bank of India, men made up the majority of respondents (67%) while women made up the remaining 33%. In contrast, 87% of responders at Punjab National Bank were men, while only 13% were women. Other survey findings showed that the majority of respondents at both banks, or 67% and 70%, respectively, belonged to the 31-45 age group and had postgraduate degrees. Additionally, it was shown that most responders were married. Additionally, the poll revealed that 93% of respondents at State Bank of India chose "Prudential Limits" as a key tool for credit risk management, whereas 40% of respondents at Punjab National Bank preferred risk rating as a tool for "Credit Risk Management." According to survey data, 37% of respondents from State Bank of India and 70% of respondents from Punjab National Bank do risk rating exercises and have begun producing MIS. In the majority of banks, the rating is displayed as "Alphabets." According to State Bank India (77%) and

Punjab National Bank (57%) the most important method for managing credit risk was "Portfolio Quality". The study's survey further revealed that the most important criteria assessed for pricing credit risk in Punjab National Bank is "Future Business Potential." Both banks evaluated "Strategic Reasons" and "Perceived Value" identically. Similar to this, Punjab National Bank and State Bank of India nearly reward the usage of derivatives in banking for managing credit risk. The conclusion was presented objectively overall.

There was a succinct study summary given. In the last ten years, India's banking industry has grown exponentially, offering a wide range of services to the nation's urban, rural, and metropolitan areas. A significant share of expenditure and income are concurrent with deposit mobilisation and credit deployment, which are at the centre of banking activity. Therefore, before considering a proposition, it is absolutely necessary to complete a credit inquiry.

Not all of the chapters had conclusions, but some of them did. A segment was also devoted to the study's general conclusion. The overall finding of this study was that regardless of the industry or size of the bank, India's credit risk control structure is on the right track and is wholly based on the RBI's published standards in this area. The prudential limitations, effective credit administration, and loan reviews are also highly critical instruments of credit risk management, even though "risk rating" is the most important one.

In light of the study's findings, numerous recommendations were made. Since the majority of respondents were not aware of the credit risk management procedures and framework to be adopted and followed at State Bank of India residential branches, one of the recommendations was that actions be taken to educate State Bank of India staff about credit risk management. Another proposal was that banks and securities businesses should properly be audited with regard to their internal controls and risk exposure.

This report, which outlines the various strategies employed by Punjab National Bank and State Bank of India and enumerates the most crucial elements that institutions consider when creating their credit risk management methods, was significantly needed.

The greatest risk that commercial banks worldwide face is credit risk, according to the abstract by Han. It was evident from the article that credit risk is steadily growing as a hazard, necessitating the necessary attention.

(Kun & Duo, 2014) described credit risk as the potential losses experienced by banks as a result of borrowers' failure to make payments. He also stated that there are three types of credit risk, including principal loss risk, interest loss risk, and profit loss risk. He divided credit risk into three further categories, including operating risk, market risk, and moral hazard. According to the article, operating risk is the possibility of suffering a loss due to insufficient or problematic internal processes, people, and systems, as well as external occurrences. He continued by saying that it might be separated into operation strategy risk and operation failure. Losses resulting from unfavourable fluctuations in interest rates, exchange rates, and the price of credit assets were classified as market risk. Additionally, he divided market risk into three major categories: interest rate risk, currency risk, and inflation risk. Moral hazard, according to Han, is when the principal and the agent both do activities that endanger other individuals.

The article in explaining the present situation of the Chinese credit risk management sought to provide the three stages of credit management experienced in China.

The typical credit risk management used throughout the planned economy was the first level. Han stated that credit was regulated by the government during the typical planned economy's 30 years preceding 1978. The idea of credit risk was not clearly understood by banks, and as a result, credit risk management focuses primarily on corruption, embezzlement, and other behaviours that violate financial discipline.

The second phase of the planned commodity economy was the management of credit risk. China's leadership declared it to be a planned commodity economy based on public ownership in October 1984. The credit industry quickly expanded after that. However, because banks were first unprepared to handle credit risk, they progressively became exposed, and during that time, credit risk increased. As a result, China developed the systems for deposit reserves, bad debt reserves, excess reserves, and asset liability ratios. Han went on to say that at this time, loans were frequently made without any guarantees to state-owned businesses, and along with the serious local government intervention, many of these businesses defaulted on their bank loans, resulting in a significant amount of bad loans.

Under the socialist economic system, credit risk management is the last phase. Chinese commercial banks have recently started to pay attention to consumer credit evaluation. However, they are lagging behind when it comes to measurement and management. He clarified that even while this system has resulted in significant advancements, there was still much work to be done.

Han addressed some of the current Chinese credit risk management's issues in the article. He listed a number of issues, including an unjustified credit organisational structure, government influence over commercial banks, an unjustified credit structure that increased credit risk, and inadequate credit management and oversight. He clarified that a lack of established mechanisms and communication contributes to business dishonesty and a rise in credit risk generally. He continued by pointing out that, despite a decline in China, the government's influence on credit risk is still felt in some sectors of the economy, which is detrimental to credit risk management. He added that these commercial banks do not use a consistent scoring system when making loans to borrowers. Additionally, they don't evaluate the use and repayment of loans on a regular basis.

Recently, Han presented several suggestions for improving China's credit risk management. A few measures to improve credit risk management were the improvement of the post responsibility and accountability system, optimization of the credit structure to reduce credit concentration, establishment of a trustworthy credit warning system, establishment of risk transformation, and establishment of the compensation mechanism of credit assets.

The goal of (Crook et al., 2007) studies "Recent Development in Consumer Credit Risk Assessment," was to identify and discuss the problems with customer credit risk assessment. According to the researchers, managing a potential customer's account using credit assessment tools from the pre-screening phases until the account is written off constitutes consumer credit risk assessment. The researchers believed that the logistic regression model was appropriate for determining client credit risk, despite the fact that many research works sought to use other models in their credit risk assessment.

In order for the lender to make an informed decision about whether to mail information about credit card and loan facilities to current and new customers, nearly every adult in the US and the UK gets assessed numerous times. Application scoring was developed to help lenders distinguish between applications they are reasonably confident would repay their credit or loan capacity. The lender often considers a sample of loan applications and carefully evaluates how well the borrowers repay their loans.

The researchers took into account behavioural scoring in addition to consumer credit scoring. Consumer credit scores are used to determine who should receive a direct mail piece. It has also been noted that government organisations, particularly tax collection agencies, utilise them to determine who is more likely to pay a tax debt. The behavioural scoring takes into account a similar premise to the consumer scoring, but the analyst now has data on the behaviours of current borrowers. The passage of the Basel II agreement was found to have the greatest impact on the methods used to determine consumer credit scores. This has had a considerable impact on how banks in developed nations estimate their reserve capital. The amount to be decided is based on the most recent scoring models, which calculate the amount due at default and the likelihood of default.

The study found that the lender uses a strategy to divide the groupings that make up the population of applicants in an effort to distinguish between applications by scoring. The most typical method for determining a significant classification on credit applications was found to be logistic regression with maximum likelihood. Applications can be categorised as good or bad using the Chi-square test, information statistics, and Sommer's D concordance statistic. The researchers studied the Kolmogorov-Smirnov (KS) statistic, the Gini coefficient, and the Receiver Operating Curve in their quest to derive important statistical models to categorise apps as good or bad (ROC). The best way to assess separation was found to be the KS statistic. The Gini coefficient between scoring models can be compared to show which creates better separation. The study found that the Gini coefficient is more informative than the ROC if the interest is on the predicting performances. The ROC is a graph that compares the percentage of good that is classified as bad against the percentage of bad that is labelled as good at all values. Since all potential cut-offs are taken into account, the ROC provides an overview of the scorecard's prejudice.

One of the most productive uses of statistical and operational research ideas has been in the area of credit scoring and risk assessment, which has significant social ramifications. Credit scoring was developed to reduce borrowing costs and to make the application process for credit cards and loans easier. The problem of finding and distinguishing between excellent and bad applications using credit scoring has been identified and solved by research.

Optimal Cut-Off Point

According to (Yin & Tian, 2014), both sensitivity and specificity are well-known metrics that depend on cut-off points. The authors explored methods to find the best cut-off point and further iterated that because the cut-off point is largely unknown, it is necessary to use optimization criteria such the Youden index to find the best cut-off point. Using the Youden index, it is possible to estimate the ideal cut-off point based on desired samples; as a result, "sensitivity and specificity at the cut-off point are associated." The ideal cut-off point for a model is found at the point where the highest Sensitivity and Specificity is reached, according to (Youden, 1950).

Similar to this, (Soureshjani & Kimiagari, 2012) investigated the best methods of determining the optimal cut-off point by utilizing logistic regression and neural network on credit scoring problem to 1,000 legitimate customers of a commercial bank. Their research showed that the appropriate cut-off amount threshold for each bank depends on the overall degree of model-generated inaccuracy. The cut-off point with the logistic regression model is "the point that the decision maker decides whether to accept the loan application or not," according to the authors, who iterated the various methods of determining the cut-off point. Customer will fall into "bad customers" category if probability exceeds cut-off point; else, they will fall into "good customers" category. The authors further suggested that, on the other hand, in order to determine the cut-off point using a neural network, "we should utilise a threshold of correct prediction on wrong one which is similar to the cut-off point in logit modelling." The use of a neural network demonstrates that a logit model's best cutoff point also works well with a neural network.

Similar to this, (Samreen and Zaidi, 2012) in their effort to create a credit scoring model for commercial banks by predicting the creditworthiness of individual borrowers, was successful in creating a model known as the Credit Scoring Model for Individuals (CSMI). As opposed to Logistic Regression (LR) and Discriminant Analysis, the findings indicated that the suggested model, "CSMI," had a higher accuracy rate with no errors (DA).

Important demographic parameters, including age, occupation, length of loan, and credit history, were taken into account by the authors in their analysis. A 0.5 cut-off point was produced by the CSMI model as the ideal value. In a sample of 250 loan applications, 96 candidates, or 38.4% of the population, were predicted to be problematic borrowers or loan defaulters, and these defaulter applicants had credit scores below the cut-off number. Additionally, it was assumed that the 154 applicants with credit scores above the cut-off point and 61.6% of the overall population would be good clients with strong creditworthiness and low default risk.

(Oliver & Thomas, 2009) researched the optimal credit score cut-offs and the pricing of regulatory capital in retail credit portfolios. Their research explored the various adjustments that must be made concerning the decision to accept or fail a loan application based on financial regulations imposed on financial institutions. The research deduced that introducing regulatory requirements such as Basel I and II will likely increase the credit score cut-off point at the expense of expected profit margin, portfolio size and return on equity (ROE).

(Stein, 2004) looked at quantitative techniques to determine the best credit cut-off point and found that a simple cut-off procedure could lead to an advanced pricing approach that is more profitable and flexible. The study also demonstrated that better credit cut-off models outperform weaker ones in terms of accuracy and success. Stein contends that the research model and the lender's cost function can be used to determine the appropriate cut-off point. The cut-off point reduces costs compared to all other possible cut-off positions. According to Stein, "Bankers frequently request a specific regulation to specify a cut-off below which credit will be denied and above which it will be provided".

Techniques Used to Handle Imbalanced Data

Imbalanced data is a common challenge in various fields such as healthcare, social sciences and finance, where the classes of interest are not evenly distributed. Imbalanced data is often characterized by many instances in one class and a relatively small number in another. This imbalance can negatively impact the performance of machine learning models since most algorithms are designed to optimize overall accuracy, which can lead to poor performance on the minority class. Consequently, several techniques have been proposed to handle imbalanced data. The main objective of this literary critique is to present a comprehensive survey of the typical strategies employed to resolve issues related to imbalanced datasets.

Resampling Techniques:

Resampling techniques aim to balance the class distribution by modifying the training set. Generally, there are two resampling techniques used to handle imbalanced datasets: oversampling and undersampling. The former involves augmenting the instances in the minority class, whereas the latter involves a reduction in the instances belonging to the majority class.

Oversampling

Oversampling involves increasing the number of instances in the minority class. Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) is a popular oversampling method. SMOTE generates synthetic instances of the minority class by creating new instances that are linear combinations of existing minority class instances. SMOTE works by selecting an instance from the minority class and selecting one of its k-nearest neighbors. It then generates a new instance by taking a linear combination of the two instances. This process is repeated until the desired number of synthetic instances is generated. The key advantage of SMOTE is that it generates synthetic instances that are similar to the existing minority class instances, which helps to preserve the characteristics of the minority class. However, SMOTE can also result in overfitting if the synthetic instances are too similar to the existing minority class instances. Borderline-SMOTE (Han et al., 2005) is a variant of SMOTE that focuses on generating synthetic instances at the borderline between the minority and majority classes. This approach can be more effective than SMOTE in scenarios where the minority class is not clearly separable from the majority class. Safe-level-SMOTE (Bunkhumpornpat et al., 2009) is another variant of SMOTE that focuses on preserving safe minority instances while oversampling. Safe minority instances are instances that are correctly classified by the classifier. This approach can be useful in scenarios where there are important minority instances that should not be oversampled. Another oversampling technique is ADASYN (Adaptive Synthetic Sampling), proposed by (He et al., 2008), which generates synthetic instances based on the density distribution of the minority class. Other oversampling techniques include Safe-Level-SMOTE (Bunkhumpornpat et al., 2009), and MWMOTE (Barua et al., 2014), among others.

Undersampling

Undersampling involves reducing the number of instances in the majority class. Exploratory undersampling (Liu et al., 2009) is a popular undersampling technique. It selects a subset of the majority class instances that are closest to the minority class instances and removes them. This approach can be effective in scenarios where the majority class instances are very different from the minority class instances. Instance hardness threshold undersampling (IHTS) (Fernández et al., 2013) is another undersampling technique that focuses on selecting instances that are difficult to classify. This approach can be useful in scenarios where the majority class instances are difficult to separate from the minority class instances.

Hybrid techniques

Hybrid techniques combine oversampling and undersampling methods. Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors (SMOTEENN) (Batista et al., 2004) is a popular hybrid technique. SMOTEENN first oversamples the minority class using SMOTE and then applies the edited nearest neighbor (ENN) algorithm to remove noisy instances from both the minority and majority classes. The ENN algorithm selects instances that have different class labels from their k nearest neighbors and removes them. SMOTEENN (Chawla et al., 2003) is a combination technique that applies SMOTE to generate synthetic instances of the minority class and then applies Tomek links to remove instances that are close to each other in the feature space. SMOTETomek (Batista et al., 2004) is another combination technique that applies SMOTE and Tomek links in sequence. Other combination techniques include G-SMOTE (Douzas & Bacao, 2019), CCR (García et al., 2012), and RUSBoost (Seiffert et al., 2010), among others.

Resampling techniques have been shown to be effective in handling imbalanced data, however, the choice of the resampling technique and its parameters may depend on the characteristics of the data, the learning algorithm used, and the evaluation metrics. The performance of the resampling techniques can also be affected by the level of imbalance in the data, the quality of the data, and the presence of noise and outliers.

In addition to the resampling techniques discussed above, some studies have proposed hybrid techniques that combine resampling with other techniques, such as feature selection, data cleaning, and data augmentation (Sun et al., 2016; Ali et al., 2020). Hybrid techniques can be more effective in handling imbalanced data than using resampling techniques alone.

Moreover, some studies have proposed deep learning models that can handle imbalanced data without the need for resampling (Jiang et al., 2018; Tan et al., 2020). Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can automatically learn features that are discriminative for the minority class. However, deep learning models may require a large amount of labelled data and computational resources.

While resampling techniques have been widely used to handle imbalanced data, they are not without their limitations. Oversampling techniques can lead to overfitting, especially when generating synthetic data, as they may introduce noise and redundancy in the training set (Bunkhumpornpat et al., 2009). Undersampling techniques may lead to information loss and reduce the representativeness of the majority class (Chawla et al., 2002). Additionally, some undersampling techniques, such as Tomek links and ENN, can be computationally expensive and may not work well with high-dimensional data (Liu et al., 2009).

Cost-Sensitive Learning:

Cost-sensitive learning is an approach where the cost of misclassification is incorporated into the learning algorithm. The cost of misclassification can be defined explicitly or implicitly. In explicit cost-sensitive learning, the cost matrix is given as input to the algorithm, and the objective function is modified to reflect the cost. In implicit cost-sensitive learning, the cost is learned from the data.

Cost-sensitive decision trees (Elkan, 2001) is an example of explicit cost-sensitive learning. In cost-sensitive decision trees, the cost matrix is incorporated into the splitting criterion. Cost-sensitive support vector machines (Lin et al., 2003) is another example of explicit cost-sensitive learning. In cost-sensitive support vector machines, the cost is incorporated into the objective function. Cost-sensitive neural networks (Zadrozny & Elkan, 2002) is an example of implicit cost-sensitive learning. In cost-sensitive neural networks, the cost is learned from the data by assigning different weights to the instances.

Cost-sensitive learning has been shown to be effective in handling imbalanced data, but it also has some limitations. Explicit cost-sensitive learning requires the cost matrix to be specified in advance, which can be difficult when the cost of misclassification is unknown or difficult to quantify (Sun et al., 2011). Implicit cost-sensitive learning, on the other hand, may be sensitive to the choice of the cost function and may require careful tuning of the learning algorithm (Zadrozny & Elkan, 2001). Additionally, cost-sensitive learning may not work well with highly imbalanced data, where the number of minority class instances is very small (Liu et al., 2009).

Ensemble Methods:

Ensemble methods are a family of techniques that combine multiple classifiers to improve the performance of the final classifier (Polikar, 2006). Bagging, boosting, and stacking are three popular ensemble methods that can be used in handling imbalanced data (He et al., 2008; Liu et al., 2009; Bensusan et al., 2000).

Bootstrap aggregating (bagging)

Bootstrap Aggregating (bagging) (Breiman, 1996) is an ensemble method that involves training multiple classifiers on different subsets of the training data. The subsets are created by randomly sampling the training data with replacement, so each subset may contain some duplicate instances and may miss some instances. Each classifier is then trained on one of these subsets, resulting in a set of diverse classifiers. During prediction, the final classification is obtained by aggregating the predictions of all classifiers, either by taking the majority vote (for classification problems) or the average (for regression problems). One popular variant of bagging is Random Subspace Method (RSM), which was introduced by (Ho, 1998). In this method, a random subset of features is selected for each classifier, and the classifiers are trained on the instances that correspond to these features. This method can be effective in handling high-dimensional data where there are many irrelevant features.

Boosting

Boosting (Freund & Schapire, 1997) is an ensemble method that involves training multiple classifiers sequentially, where each classifier is trained on a modified version of the training data. During training, more emphasis is given to misclassified instances, so the subsequent classifiers will try to improve the accuracy of these instances. AdaBoost (Zadrozny & Elkan, 2001) is a popular boosting algorithm that assigns higher weights to misclassified instances in each subsequent iteration. In this method, the instances are assigned weights, and the classifiers are trained on the weighted data. The weights are updated after each iteration, so the subsequent classifiers will focus more on misclassified instances. During prediction, the final classification is obtained by aggregating the predictions of all classifiers, where the classifiers that perform well have higher weights.

Stacking

Stacking is an ensemble method that involves training multiple base classifiers of different types and combining their predictions using a meta-learner. The meta-learner takes the predictions of the base classifiers as input and outputs the final prediction. The base classifiers can be diverse in terms of their learning algorithms, features, and hyperparameters. This diversity can help capture different aspects of the data and improve the overall accuracy of the model. During training, the base classifiers are trained on the training data, and their predictions are used to train the meta-learner. During prediction, the base classifiers make predictions on the test data, and their predictions are used as input to the meta-learner to make the final prediction. Stacking can be used to combine the strengths of different classifiers and can improve the accuracy of the model. However, it can be computationally expensive and may require a large number of base classifiers to achieve good performance. One popular variant of stacking is Hierarchical Stacking, which was introduced by (Bensusan et al., 2000). In this method, the base classifiers are organized in a hierarchical structure, where each level corresponds to a different level of abstraction of the data. The predictions of the base classifiers at each level are combined using a meta-learner, which outputs the final prediction. Another variant of stacking is Blending, which was introduced by (Wolpert, 1992). In this method, the training data is split into two parts: the first part is used to train the base classifiers, and the second part is used to train the meta-learner. During prediction, the base classifiers make predictions on the test data, and their predictions are used as input to the meta-learner to make the final prediction.

There are several variants of each ensemble method. For example, Random Subspace Method is a variant of bagging that trains multiple classifiers on different subsets of the features instead of the instances (Ho, 1998). ADASYN-Boost is a variant of boosting that uses a synthetic minority oversampling technique to generate new instances of the minority class (He et al., 2008). Hierarchical Stacking is a variant of stacking that organizes the base classifiers in a hierarchical structure based on the abstraction level of the data (Bensusan et al., 2000).

Ensemble methods have been shown to be effective in handling imbalanced data, but they may also have some limitations. Bagging and boosting may not work well with highly imbalanced data, where the number of minority class instances is very small (Liu et al., 2009). Additionally, ensemble methods may be computationally expensive and may require careful tuning of the parameters (He & Garcia, 2009).

One-Class Classification:

OCC is a type of unsupervised learning where the goal is to identify and classify instances from one class of interest. The technique assumes that the data is skewed towards the majority class, and therefore, the instances from the minority class are few and may be difficult to detect. OCC aims to identify these instances and model them as a separate class.

OCC has been applied in various domains to handle imbalanced data. In fraud detection, OCC has been used to identify unusual financial transactions, such as credit card fraud (Chandola, Banerjee, & Kumar, 2009). In medical diagnosis, OCC has been applied to detect rare diseases that are difficult to diagnose (Witten & Frank, 2005). In anomaly detection, OCC has been used to identify unusual patterns in network traffic (Patcha & Park, 2007). These applications show that OCC can be used in various fields to address the challenge of imbalanced data.

Several techniques have been proposed for OCC, including support vector data description (SVDD), one-class random forest (OCRF), and autoencoder-based OCC. SVDD is a technique that models the target class as a hypersphere in the feature space (Tax & Duin, 1999). The goal is to find the smallest hypersphere that can contain all the instances from the target class. The instances outside the hypersphere are considered outliers. OCRF is a modification of the traditional random forest algorithm, where the model is trained on the minority class only (Breiman, 2001). The model consists of decision trees that are trained on random subsets of the minority class. Autoencoder-based OCC is a deep learning technique that uses a neural network to learn a compressed representation of the data (Andrews et al., 2016). The model is trained on the minority class only, and the goal is to learn a compact representation that can capture the features of the minority class.

One-class classification has been shown to be effective in handling imbalanced data, but it also has some limitations. One-class classification techniques may not work well when the minority class is not well-defined or is difficult to characterize (He & Garcia, 2009). Additionally, one-class classification may be sensitive to the choice of the kernel function and may require careful tuning of the parameters (Tax & Duin, 2004).

Handling imbalanced data is a common challenge in various fields, and the performance of machine learning models can be negatively impacted by class imbalance. Resampling techniques, including oversampling, undersampling, and hybrid techniques, have been proposed to handle imbalanced data. These techniques aim to balance the class distribution by modifying the training set. Resampling techniques have been shown to be effective, but the choice of the technique and its parameters may depend on various factors such as the characteristics of the data and the learning algorithm used. In addition to resampling techniques, hybrid techniques that combine resampling with other techniques have been proposed, which can be more effective in handling imbalanced data. Furthermore, some studies have proposed deep learning models that can handle imbalanced data without the need for resampling. Overall, the choice of the technique depends on the specific problem and available resources, and the performance of the techniques can be influenced by various factors.

Imbalanced data is often dealt with by using different techniques, such as resampling, ensemble methods, or one-class classification, all of which require manipulating the data in some way. However, this research posits that no such manipulation is necessary. Instead, the proposed methodology involves utilizing sensitivity and specificity plots to identify an optimal cut-off point for classifying the data. This approach not only avoids the potential biases introduced by data manipulation but also offers a more robust and reliable classification strategy.

Chapter Summary

None of these statistical methods were found to be the best method to use for credit scoring, despite the fact that the majority of the articles and research works that were critiqued as part of this review focused on the use of credit risk analysis techniques by way of regression analysis, logistic regression analysis, correlation analysis, optimization models, descriptive statistics, as well as the use of well-structured credit scoring models and accounting models. This is corroborated by (Abdou & Pointon, 2011), who found that there is no one best statistical technique used in developing credit scoring models and that the best technique for all circumstances does not yet exist. They reached this conclusion after carefully examining 214 articles, books, and theses that dealt with credit scoring applications in various fields, particularly in finance and banking. However, due to its assumptions and simplicity in interpretation, logistic regression is commonly used. Additionally, between 60% and 80% accuracy is often produced by logistic regression models. (Le Bellac & Viricel, 2016; Einarsson, 2008).

Additionally, the literature review has revealed that there is lack of further analysis on specificity and sensitivity to identify the optimal cut-off point of credit scoring models developed.

Therefore, the main goal of this research project is to create a binary regression model that can distinguish between Ghanaian commercial bank loan defaulters and non-defaulters. The study will subsequently determine the optimal cut-off point for the bank's loan decision-making and posit that the optimal cut-off point may be used as an alternative to utilizing resampling techniques to deal with imbalanced data typical of financial data. This study will further investigate the implications of altering the cut-off point on the commercial bank's choice and the derived model.

CHAPTER THREE

METHODOLOGY

Introduction

This chapter contains the data characterization and overview of the methods used to construct the proposed models to analyze the data under study.

The population under study are borrowers i.e., individuals who have loans with a commercial bank in Ghana. The data is institutional-based obtained from one of the indigenous commercial banks in Ghana (data specific) because the information required to qualify for a loan differ from bank to bank although the data obtained from the bank may be representative of the population of borrowers in Ghana due to the fact that all banks in Ghana use the regulator's minimum Know Your Customer (KYC) Policy as prescribed by the regulator, Bank of Ghana (Bank of Ghana, n.d.; BOG & FIC, 2018). Furthermore, other information on borrowers, in addition to the minimum KYC requirements, may result in unique credit scoring models for each bank. However, the methodology described below can be applied to any data on borrowers obtained from any bank in Ghana.

The data consist of 9,590 borrowing customers. These loans are either on schedule or in default as of December 31st, 2019. In order to maintain customer confidentiality, The names, telephone numbers and residential addresses of the borrowers were removed. This resulted in the data having ten variables on each borrower; Amount financed, Tenor of the loan, Interest rate on loan, "Age of the borrower", "Gender of the borrower", "Marital status of the borrower", Type of Employment of the borrower, Loan description (or

Type), Economic sector in which the borrower operates and Loan default. Six borrowers were removed due to data entry errors (misclassification by data entry clerks). Of the remaining 9,584 borrowers, 896 had of them missing their Marital Status or Type of employment.

Additionally, one borrower had both his/her Marital Status and Type of Employment missing. Information supplied by the bank on the missingness was that these were omissions made by data entry clerks and not because of non-response. The 897 borrowers represent approximately 9.3% of the total sample size. We do not expect that deleting these 897 borrowers out of 9,584 borrowers will introduce any statistically significant bias in our data analysis (Clavel et al., 2014); hence, these borrowers will be discarded and will not be included in the analysis. Some extreme values in the amount financed column held vital information, so they were not treated as outliers. The remaining data set contains 8,687 borrowers and ten variables.

Subsequently, IBM's SPSS Modeler will be used to build 16 binary models using the algorithms listed in Table 2 below:

Table 2: List of Algorithms

S/N	Model
1.	Bayesian Network
2.	C5 (a decision tree algorithm)
3.	Chi-squared Automatic Interaction Detection (CHAID)
4.	Classification and Regression Tree (C&R)
5.	Decision List
6.	Discriminant
7.	K-Nearest Neighbors (KNN)
8.	Linear Support Vector Machine (LSVM)
9.	Logistic Regression
10.	Neural Net
11.	Quick, Unbiased, Efficient Statistical Trees (QUEST)
12.	Random Forest
13.	Random Trees
14.	Support Vector Machine (SVM)
15.	Tree-AS
16.	XGBoost Linear
17.	XGBoost Tree

Author's Construction

The algorithms will then be grouped into five groups in accordance to similar underlying methodology as shown below in Table 3 below:

Table 3: Grouped Algorithms

1. Tree-based algorithms:	2. Linear algorithms:	3. Non-linear algorithms:
C&R Tree	Discriminant Analysis	Bayesian Network
C5	Logistic Regression	Neural Network
CHAID	LSVM	SVM
Random Forest	XGBoost Linear	
Random Trees		
Tree-AS		
XGBoost Tree		
4. Rule-based algorithms:	**5. Other**	
Decision List	K-NN	
QUEST		

Author's Construction

The dataset will be divided into two parts: the train sample and the test sample, with a ratio of 70% to 30%, respectively. The train sample is a randomly selected 70% subset of the data that follows the Bernoulli distribution and will be used to train the models. The remaining 30% of the data will be used to validate the models' performance. Although there is no definitive formula in the literature regarding the optimal ratio of data partitioning, a 70:30 ratio is commonly used (Xu & Goodacre, 2018).

The logistic regression model, as well as 16 other models outlined in Table 2, will be developed using the 70% train sample. The resulting models will be then applied to the 30% test sample, and the area under the curve (AUC) of each model will be compared. The top-performing model in each group, as specified in Table 3, will be selected. The AUC values of the best models in each group will then be compared to the logistic regression model developed using the test sample.

Subsequently, resampling techniques such as oversampling and undersampling will be applied to the data, and each data set will be divided into 70% for training the models and 30% for testing the models as earlier specified. The AUC values of each model on the test sample in each group will then be compared to that of the logistic regression model's performance at the derived optimal cut-off point.

Data Characterization

Loan default is the dependent variable where we would like to predict whether a borrower will default or not default on a loan using the other nine (9) independent variables. The variables are summarized in Table 4 and Table 5 below.

Table 4. Characteristics of the Categorical Variables

Variable	Description	Values
LOANDEF	Loan Default (describes a borrower as defaulting or otherwise)	0 = No 1 = Yes
GENDER	Gender of the borrower	1 = Female 2 = Male
MARSTAT	Marital Status of the borrower	1 = Divorced 2 = Engaged 3 = Married 4 = Single 5 = Widowed
EMPTYP	This represents the employment status of the borrower	1 = Full Time 2 = Part Time 3 = Retired 4 = Other
LOANDESC	This describes the type of loan given to the customer	1 = CAGD Salary Loan Scheme 2 = Commercial Loans 3 = Institutional Workers Loan Scheme 4 = Microcredit Loan Sheme 5 = Salaried Workers Loan Scheme
ECOSECT	This denotes the economic sector in which the borrower works or operates	1 = Primary 2 = Secondary 3 = Tertiary

Author's Construction

A borrower who has concurrently defaulted a month or more as of the end of December 31st, 2019, is classified as default; however, a borrower who had defaulted a month or more during the tenor of the loan but has paid up all outstanding months on or before December 31st, 2019, is on schedule and hence is classified as non-default.

The "Other" classification in Type of Employment represents other employment statuses that are not classified under the mentioned categories, e.g., Consultants, Voluntary Workers, Interns, NSS etc.

CAGD Salary Loans are loans given to all borrowers who receive their salary from the Controller and Accountant General's Department. Commercial Loans are loans granted to borrowers specifically for commercial activities, and the source of repayment is from the same commercial activities. Institutional/University Workers Loans are loans granted to university staff. Micro Credit Loans are loans given to borrowers in the economy's informal sector, otherwise known as the "tabletop" retailers. Salary Workers Loans are loans given to any other salaried worker who receives a salary from a corporate body but not CAGD or a University.

Borrowers are categorised in the Economic Sector according to the three-sector economics model. As a result, economies are divided into three main areas of activity: "primary", which focuses on the extraction of raw materials, "secondary", which focuses on manufacturing; and tertiary, which focuses on service industries, which are necessary to support the transportation, distribution, and sale of goods made in the secondary sector.

Table 5. Characteristics of the Numerical Variables

Variable	Description	Values
AMTFIN	Amount Financed variable denotes the amount of money credited or given to the borrower as a loan.	Range from GH¢550.59 to GH¢470,175.44.
TENOR	The Tenor represents the life span or time frame of the loan or the period in which the loan should be repaid in days. In other words, the length of time until the total amount of the loan is due.	Range between 30 days (a month) to 2,554 days (7 years) inclusive.
INTRATE	Interest Rate is the amount charged by the bank to borrow its money. Sometimes referred to as the Lending Rate; it is expressed as an annual percentage on the amount financed or principal.	Range from 20% to 46% per annum.
AGE	This denotes the age of the borrower at the time of loan contract initiation.	Range from 20 to 83 years.

Author's Construction

Discriminant analyses and linear regression can be applied to data of the above-described data characteristics; however, they are limited by the assumptions of normality. Here, we have the dependent variable being binary and more than one of the independent variables being categorical. The known and the most popular statistical technique to apply to this type of data that does not consider normality is the logistic regression model.

In comparison to discriminant analysis and linear regression, logistic regression employs maximum likelihood estimators, which involve computationally more difficult techniques. However, the output of logistic regression models is a percentage term that is directly explicable and may be utilised to carry out operational actions like setting cut-off points or values because they are not restricted by the assumption of normality. These benefits make the logit model the ideal choice for our analysis.

Logistic Regression

It may have been noticed from our data description that this is a multivariate case; hence, multiple logistic regression methodology will be applied.

Multiple logistic regression model (MLRM)

As we already stated above, we consider a collection of p independent variables denoted by the vector $\boldsymbol{x}' = \left(x_1, x_2, \ldots, x_p\right)$. Let us use the following notation to represent the conditional probability that the outcome is true: $\Pr(Y = 1|\mathbf{x}) = \pi(\boldsymbol{x})$. The logit of the multiple logistic regression model is given by the equation

$$g(\boldsymbol{x}) = ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{1}$$

where for the multiple logistic regression model,

$$\pi(\boldsymbol{x}) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{2}$$

In our case where we have some of the variables being in nominal scale, we know in general that if a nominal scaled variable has k possible values, then, $k - 1$ design variables are needed. The reason for using one less than the number of values is that, unless stated otherwise, our models have a constant term. To illustrate the notation used for design variables in this work, suppose that the jth independent variable x_j has k_j levels. The $k_j - 1$ design variables will be denoted by D_{jl} and the coefficients for the design variables will be denoted by β_{jl}, where $l = 1, 2, \ldots, k_j - 1$. Hence, the logit for a model with p variables, with jth variables being discrete is

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \sum_{i=1}^{k_j - 1} \beta_{jl} D_{jl} + \beta_p x_p \tag{3}$$

Fitting the MLRM

Assume that our sample n contains a variety of impartial observations $(\boldsymbol{x}_i, y_i)$, where $i = 1, 2, \ldots, n$. To fit the model, we must determine the vector's estimations $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$. We use the maximum likelihood function for the estimation. Thus, for those pairs $(\boldsymbol{x}_i, y_i)$, where $y_i = 1$, the contribution to the likelihood function is $\pi(\boldsymbol{x}_i)$, and for those pairs where $y_i = 0$ the contribution to the likelihood function is $1 - \pi(\boldsymbol{x}_i)$. We conveniently express the contribution to the likelihood function for the pair $(\boldsymbol{x}_i, y_i)$ as

$$\pi(\boldsymbol{x}_i)^{y_i}[1 - \pi(\boldsymbol{x}_i)]^{1-y_i} \tag{4}$$

Given that it is assumed that each observation is independent, the likelihood function is obtained as the product of the terms given in Equation (4) as follows

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}[1 - \pi(\boldsymbol{x}_i)]^{1-y_i} \tag{5}$$

According to the principle of maximum probability, we should estimate values based on those that maximise the expression in Equation (5). However, it is easier mathematically to work with the log of Equation (5). The loglikelihood is defined as

$$L(\boldsymbol{\beta}) = ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \, ln[\pi(\boldsymbol{x}_i)] + (1 - y_i) \, ln[1 - \pi(\boldsymbol{x}_i)]\} \tag{6}$$

To find the value $\boldsymbol{\beta}$ that maximizes, $L(\boldsymbol{\beta})$ we have to differentiate $L(\boldsymbol{\beta})$ with respect to $\beta_0, \beta_1, \ldots, \beta_p$. There will be $p + 1$ likelihood equations obtained by differentiating the log-likelihood function with respect to the

$p + 1$ coefficients. The likelihood equations that result may be expressed as follows

$$\sum_{i=1}^{n}[y_i - \pi(\boldsymbol{x}_i)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\boldsymbol{x}_i)] = 0$$

for $j = 1, 2, \dots, p$

Let $\widehat{\boldsymbol{\beta}}$ denote the solution to these equations. Thus, the fitted values for the multiple logistic regression model are $\hat{\pi} = (\boldsymbol{x}_i)$, the value of the expression in Equation (2) computed using $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{x}_i$.

The principle of maximum likelihood estimation is used in estimating the variances and covariances of the computed coefficients. This theory implies that the estimators are contained in the second partial derivatives matrix of the log-likelihood function. These partial derivatives have the following general form.

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = \sum_{i=1}^{n} x_{ij}^2\, \pi_i(1 - \pi_i) \tag{7}$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = \sum_{i=1}^{n} x_{ij} x_{il}\, \pi_i(1 - \pi_i) \tag{8}$$

for $j = 0, 1, 2, \dots, p$ where π_i denotes $\pi(\boldsymbol{x}_i)$.

Let the $(p + 1) \times (p + 1)$ matrix containing the negative of the terms given in equation (7) and (8) be denoted as $\boldsymbol{I}(\boldsymbol{\beta})$. The variances and covariance's of the estimated coefficients are obtained from the inverse of the observed information matrix, $\boldsymbol{I}(\boldsymbol{\beta})$, denoted as$\text{Var}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$We denote $\text{Var}(\beta_j)$ to be the jth diagonal elements of this matrix, "which is a variance of $\hat{\beta}_j$, and $\text{Cov}(\beta_j, \beta_l)$ to denote an arbitrary off-diagonal element, which is the covariance of $\hat{\beta}_j$ and $\hat{\beta}_l$. The estimators of the variances and covariances, which will be denoted by, $V\hat{a}r(\widehat{\boldsymbol{\beta}})$ are obtained by evaluating $\text{Var}(\boldsymbol{\beta})$ at $\widehat{\boldsymbol{\beta}}$. We use $V\hat{a}r(\hat{\beta}_j)$ and $C\hat{o}v(\hat{\beta}_j, \hat{\beta}_l)$, where both$j, l = 0, 1, 2, \ldots, p$ to denote values in this matrix. For the most part, we only use the estimated standard errors of the estimated coefficients, which we denote as $\widehat{\text{SE}} = (\hat{\beta}_j) = [V\hat{a}r(\hat{\beta}_j)]^{\frac{1}{2}}$ for $j = 0, 1, 2, \ldots, p$

This notation is used to create techniques for confidence interval estimates and coefficient testing.

A formulation of the information matrix that is useful when discussing model fitting and assessment of fit is $\hat{\boldsymbol{I}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}'\widehat{\boldsymbol{V}}\boldsymbol{X}$ where $\boldsymbol{X}$ is an $n \times (p+1)$ matrix containing the data for each subject and $\boldsymbol{V}$ is an $n \times n$ diagonal matrix with general element $\hat{\pi}_j(1 - \hat{\pi}_j)$ i.e., the matrix $\boldsymbol{X}$is

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

and the matrix $\boldsymbol{V}$ is

$$\boldsymbol{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

where $\hat{\pi}_i = \hat{\pi}(\boldsymbol{x}_i)$ are value of Equation (2) using $\widehat{\boldsymbol{\beta}}$ and the covariates of the subject i, $\boldsymbol{x}_i$.

Testing for the significance of the model

We start the model evaluation phase after fitting the MLRM. First, we evaluate the significance of the model's variables. The log-likelihood function stated in Equation (6) compares observed and predicted values.

The likelihood function is used to compare observed and predicted values and is based on the following expression.

$$D = -2\,ln\left[\frac{(\text{likelihood of the fitted model})}{(\text{likelihood of the saturated model})}\right] \tag{9}$$

Using equation (6), equation (9) above called the Deviance becomes

$$D = -2\sum_{i=1}^{n}\left[y_i\,ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\,ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \tag{10}$$

where $\hat{\pi}_i = \hat{\pi}(\boldsymbol{x}_i)$.

In our case, where the outcome variable values are either 0 or 1, the likelihood of the saturated model is identically equal to 1. It follows from the definition of a saturated model that $\hat{\pi}_i = y_i$ and the likelihood is

$$l(\text{saturated model}) = \prod_{i=1}^{n} y_i^{y_i} \times (1-y_i)^{(1-y_i)} = 1$$

Thus, it follows from Equation (9) that the deviance is

$$D = -2\,ln(\text{likelihood of the fitted model})$$

We evaluate the result of the equation with and without the independent variable to determine the relevance of the independent variable. The modification brought about by the independent variable's inclusion in the model is

$$G = D(\text{model without the variable}) - D(\text{model with the variable})$$

Because the likelihood of the saturated model is always common to both values of D being differenced, G can be expressed as

$$G = -2\, ln\left[\frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})}\right]$$

This is the likelihood ratio test for the overall significance of the p coefficients for the independent variables in the model.

Under the hypothesis that each $p+1$ coefficient is equal to zero, the statistic, G, follows a chi-square distribution with $p+1$ degrees of freedom.

Two alternative tests can be used to evaluate the model's significance that is comparable to the likelihood ratio test. Both the Score test and the Wald test.

The following vector-matrix calculation yields the multivariable equivalent of the Wald test.

$$W = \widehat{\boldsymbol{\beta}}'\left[V\hat{a}r(\widehat{\boldsymbol{\beta}})\right]^{-1}\widehat{\boldsymbol{\beta}}$$

$$= \widehat{\boldsymbol{\beta}}'(\boldsymbol{X}\widehat{\boldsymbol{V}}\boldsymbol{X})\widehat{\boldsymbol{\beta}}$$

which is distributed as chi-square with $p+1$ degrees of freedom under the hypothesis that each of the $p+1$ coefficients is equal to zero. The multivariable Wald test, equivalent to the likelihood ratio test for the significance of the fitted model, is based on just the p slope coefficients and is obtained by eliminating $\hat{\beta}_0$ from $\widehat{\boldsymbol{\beta}}$ and the relevant row (first or last) and column (first or last) from $(\boldsymbol{X}'\widehat{\boldsymbol{V}}\boldsymbol{X})$

The multivariable analogue of the Score test for the significance of the model is based on the distribution of the p derivatives of $L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ The computation for this test is on par with the Wald test in terms of complexity. In order to define it in depth, extra notation would have to be introduced, which would be of little benefit to the rest of this work. As a result, we direct the reader to (Cox & Hinkley, 1974; Dobson, 2002).

The likelihood ratio test will be applied in our research. This is due to the fact that, although being the simplest of the three tests, the Wald test appears to be the weakest (Hosmer et al., 2013; Moore & Rao, 1967). Wald technique must always give symmetric confidence intervals regarding the MLE because it is unable to capture asymmetry in the likelihood function. This limitation does not apply to the likelihood ratio (Hosmer et al., 2013).

Confidence interval estimation

The general expression for the estimator of the logit for a model containing p covariates is

$$\hat{g}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p \tag{11}$$

An alternative way to express the estimator of the logit in Equation (11) is through the use of vector notation $\hat{g}(\boldsymbol{x}) = \boldsymbol{x}'\widehat{\boldsymbol{\beta}}$, where the vector $\widehat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ denotes the estimator of the $p + 1$ coefficients and the vector $\boldsymbol{x}' = (x_1, x_2, x_3, \ldots, x_p)$ represents a set of values of the p −covariates in the model and the constant, $x_0 = 1$.

An expression for the estimator of the variance of the estimator of the logit in Equation (11) is

$$V\hat{a}r[\hat{g}(\boldsymbol{x})] = \sum_{j=o}^{p} x_j^2 \, V\hat{a}r(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k C\hat{o}v(\hat{\beta}_j, \hat{\beta}_k) \tag{12}$$

By utilizing the matrix equation for the estimator of the variance of the estimator of the coefficients, we can express this result much more succinctly. From the observed information matrix's expression, we have that

$$V\hat{a}r[\widehat{\boldsymbol{\beta}}] = \left(\boldsymbol{X}'\widehat{\boldsymbol{V}}\boldsymbol{X}\right)^{-1} \tag{13}$$

It follows from Equation (13) that an equivalent expression for the estimator in Equation (12) is

$$V\hat{a}r[\hat{g}(\boldsymbol{x})] = \boldsymbol{x}'V\hat{a}r(\widehat{\boldsymbol{\beta}})\boldsymbol{x}$$

$$= \boldsymbol{x}'(\boldsymbol{X}'\widehat{\boldsymbol{V}}\boldsymbol{X})^{-1}\boldsymbol{x} \quad (14)$$

The statistical software SPSS employed for this study allows the user to choose to add a new variable to the SPSS logistic regression software program that contains the estimated values of Equation (14), or the standard error for all observed values of the covariates for each individual in the data set. This innovation also enables users to execute routine calculations of fitted values and confidence interval estimates, hence reducing computing burden on Equation (14)**'s** matrix calculations.

Measures of goodness of fit

The Goodness of Fit is assessed over the fitted values determined by the covariates in the $\boldsymbol{x}$model. For instance, in our work, suppose that our fitted model contains p independent variables, and let j denote the number of distinct values of $\boldsymbol{x}$ observed. If some subjects have the same value of $\boldsymbol{x}$ then $J < n$. We denote the number of subjects $\boldsymbol{x}' = (x_1, x_2, x_3, \ldots, x_p)$ with $\boldsymbol{x} = \boldsymbol{x}_j$ by m_j,$j = 1,2,3, \ldots, J$. It follows that $\sum m_j = n$. Let y_j denote the number of responses, $y = 1$, among the m_j subjects with $\boldsymbol{x} = \boldsymbol{x}_j$. It follows that $\sum y_j = n_1$, the total number of subjects with $y = 1$. The statistical distributions of the summary goodness of fit statistics are obtained by letting n become large while holding the number of parameters, coefficients, in the model fixed. If the number of covariate patterns also increases with n then each value of m_j tends to be small. Distributional results obtained under the condition that only n becomes large are said to be based on n −asymptotics. If we fix $J < n$ and let n become large, then each value of m_j also tends to become large.

Distributional results based on each m_j becoming large are said to be based on m −asymptotics. We assume that $J \approx n$ since in our model there is at least one continuous covariate.

Sum-of-squares, deviation, and the Pearson chi-square statistic

We denote the fitted value for the j^{th} covariate pattern as

$$\hat{y}_j = m_j\hat{\pi}_j = m_j\left\{\frac{e^{\hat{g}(\boldsymbol{x}_j)}}{1+e^{\hat{g}(\boldsymbol{x}_j)}}\right\}$$

where $\hat{g}(\boldsymbol{x}_j) = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \hat{\beta}_2 x_{j2} + \ldots + \hat{\beta}_p x_{jp}$ is the estimated logit.

We begin by considering three measures of the difference between the observed and the fitted values: the Pearson residual, the deviance residual, and the residual used in linear regression. For a particular covariate pattern, the Pearson residual is

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}} \tag{15}$$

The summary statistic based on these residuals is the Pearson chi-square statistic

$$X^2 = \sum_{j=1}^{J}\left[r(y_j, \hat{\pi}_j)\right]^2$$

The deviance residual is

$$d(y_j, \hat{\pi}_j) = \pm\left\{2\left[y_j\, ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\, ln\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{\frac{1}{2}} \tag{16}$$

where the sign + or − is the same as the sign of $(y_j - m_j\hat{\pi}_j)$. For covariate patterns with $y_j = 0$ the deviance residual is

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j|ln(1-\hat{\pi}_j)|}$$

and the deviance residual when $y_j = m_j$ is

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j |ln(\hat{\pi}_j)|}$$

The summary statistic based on the deviance residuals is the deviance

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2$$

In a setting where$j = n$, this is the same quantity shown in equation (10).

The linear regression-like residual is defined as the difference between the observed and predicted outcome (as determined by the model), namely

$$s(y_j, \hat{\pi}_j) = (y_j - m_j \hat{\pi}_j)$$

and the fit statistic is the sum-of-squares

$$s = \sum_{j=1}^{J} s(y_j, \hat{\pi}_j)^2$$

According to (Hosmer et al., 2013), the distribution of the "statistics X^2 and D under the assumption that the fitted model is correct in all aspects, and it is supposed to be chi-square with degrees of freedom equal to $J - (p + 1)$. They stated that for the deviance this statement follows from the fact that D is the likelihood ratio test statistic of a saturated model with J parameters versus the fitted model with $p + 1$ parameters. A similar theory provides the null distribution of X^2. They explained that the problem is that when $J \approx n$, the distribution is obtained under n −asymptotics, and hence the number of parameters is increased at the same rate as the sample size. Thus, p-values calculated for these two statistics when $J \approx n$, using the $\chi^2(J - p - 1)$ distribution, are incorrect.

One way to avoid the above-noted difficulties with the distributions of X^2 and D when $J \approx n$ is to group the data in such a way that m −asymptotics can be used. The chi-square distribution is a good approximation of the Hosmer-Lemeshow goodness of fit statistic, as will be covered below.

The Hosmer–Lemeshow tests

Based on the predicted probability' values, Hosmer and Lemeshow (1980) and Lemeshow and Hosmer (1982) advocated grouping. The Hosmer-Lemeshow goodness of fit statistic, $\hat{C}$, is obtained by calculating the Pearson chi-square statistic from the $g \times 2$ table of observed and estimated expected frequencies. A formula defining the calculation of $\hat{C}$ is as follows.

$$\hat{C} = \sum_{k=1}^{g} \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right]$$

where

$$o_{1k} = \sum_{j=1}^{c_k} y_j ,$$

$$o_{0k} = \sum_{j=1}^{c_k} (m_j - y_j) ,$$

$$\hat{e}_{1k} = \sum_{j=1}^{c_k} m_j \hat{\pi}_j ,$$

$$\hat{e}_{0k} = \sum_{j=1}^{c_k} m_j (1 - \hat{\pi}_j)$$

The "c_k" is the number of covariate patterns in the kth group. It can be shown that

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(o_{1k} - n_k^{'} \overline{\pi}_k\right)^2}{n_k^{'} \overline{\pi}_k (1 - \overline{\pi}_k)}$$

where "$\overline{\pi}_k$" is the average estimated probability in the k^{th} group;

$$\bar{\pi}_k = \frac{1}{n_k'} \sum_{j=1}^{c_k} m_j \hat{\pi}_j$$

Using an extensive set of simulations, "Hosmer and Lemeshow (1980) demonstrated that, when $J = n$ and the fitted logistic regression model is the correct model, the distribution of the statistic $\hat{C}$ is well approximated by the chi-square distribution with $g - 2$ degrees of freedom$\chi^2(g - 2)$.

Logistic Regression Diagnostics

In logistic regression, there are binomial errors, and as a result, the error variance is a function of the conditional mean

$$\begin{aligned} var(Y_j|\boldsymbol{x}_j) &= m_j(Y_j|\boldsymbol{x}_j) \times [1 - (Y_j|\boldsymbol{x}_j)] \\ &= m_j\pi(x_j)[1 - \pi(x_j)] \end{aligned}$$

Let r_j and d_j denote the values of the expressions given in Equations (15) and (16), respectively, for covariate pattern x_j. Since each residual has been divided by an approximate estimate of its standard error, we expect that if the logistic regression model is correct, these quantities have a mean approximately equal to zero and a variance approximately equal to one.

In addition to the residuals for each covariate pattern, other quantities central to the formation and interpretation of logistic regression diagnostics are the "hat" matrix and its leverage values. Using weighted least squares linear regression as a model, (Pregibon, 1981) derived a linear approximation to the fitted values, yielding a hat matrix for logistic regression.

$$\boldsymbol{H} = \boldsymbol{V}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{1/2} \quad (17)$$

where $\boldsymbol{V}$ is a $j \times j$ diagonal matrix with general element$v_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)]$

Let the quantity h_j (leverage) denote the j^{th} diagonal element of the matrix $\boldsymbol{H}$ defined in the equation (17). It may be shown that

$$h_j = m_j\hat{\pi}(\boldsymbol{x}_j)\left[1-\hat{\pi}(\boldsymbol{x}_j)\right]\boldsymbol{x}_j^{'}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}_j$$

$$= v_j \times b_j$$

where $v_j = m_j\hat{\pi}(\boldsymbol{x}_j)\left[1-\hat{\pi}(\boldsymbol{x}_j)\right]$ is the model-based estimator of the variance of y_j, and $b_j = \boldsymbol{x}_j'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}_j$ is the weighted distance of $\boldsymbol{x}_j$ from $\bar{\boldsymbol{x}}$, where $\boldsymbol{x}_j' = (1, x_{1j}, x_{2j}, \ldots x_{pj})$ is the vector of covariate values defining the jth covariate pattern and $\bar{\boldsymbol{x}}$ is the vector of means. The sum of the diagonal elements of $\boldsymbol{H}$ is, $\sum_{j=1}^{j} h_j = (p + 1)$, the number of parameters in the model. When we formulate the hat matrix for logistic regression as an $n \times n$ matrix then each diagonal element is bounded from above by $\frac{1}{m_j}$, where m_j is the total number of subjects with the same covariate pattern. When the hat matrix is based upon data grouped by covariate pattern, the upper bound for any diagonal element is one.”

The Pearson residual defined in Equation (15), computed individually for each subject with this covariate pattern is

$$r_i = \frac{(0-\hat{\pi}_j)}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}}$$

$$= -\sqrt{\frac{\hat{\pi}_j}{(1-\hat{\pi}_j)}}$$

while the Pearson residual based on all subjects with this covariate pattern is

$$r_i = \frac{(0-m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

$$= -\sqrt{m_j}\sqrt{\frac{\hat{\pi}_j}{(1-\hat{\pi}_j)}}$$

which increases negatively as m_j increases.

Another useful diagnostic statistic is one that examines the effect that deleting all subjects with a particular covariate pattern has on the value of the estimated coefficients and the overall summary measures of fit X^2 and$D$. It is obtained as the standardized difference between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(-j)}$, where these represent the maximum likelihood estimates computed using all J covariate patterns and excluding the m_j subjects with pattern $\boldsymbol{x}_j$ respectively, and standardizing via the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. (Pregibon, 1980) showed, to a linear approximation, that this quantity for logistic regression is

$$\Delta\hat{\boldsymbol{\beta}}_j = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-j)})'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-j)})$$

$$= \frac{r_j^2 h_j}{(1-h_j)^2}$$

$$= \frac{r_{sj}^2 h_j}{(1-h_j)}$$

Using linear approximations, it can be shown that the decrease in the value of the Pearson chi-square statistic due to the deletion of the subjects with covariate patterns $\boldsymbol{x}_j$ is

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)} = r_{sj}^2 \quad (18)$$

A similar quantity may be obtained for the change in the deviance,

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)}$$

If we replace r_j^2 by d_j^2 , this yields the approximation

$$\Delta D_j = \frac{d_j^2}{(1-h_j)} \quad (19)$$

which is similar in form to the expression in Equation (18)

In linear regression, the diagnostics' value is frequently interpreted in two ways. The first is visual. The second assumes the fitted model is valid and uses the linear regression distribution theory to generate the diagnostic distribution. We use visual inspection largely in logistic regression since the distribution of diagnostics under the premise that the model fits is only known in a few limited instances.

Having defined diagnostic statistics, $(r_j, d_j, h_j, \Delta X_j^2, \Delta D_j, \Delta\hat{\boldsymbol{\beta}}_j)$ we will plot the following to aid in the analysis of the logistic regression diagnostics:

1. h_j versus $\hat{\pi}_j$
2. ΔD_j versus $\hat{\pi}_j$
3. $\Delta\hat{\boldsymbol{\beta}}_j$ versus $\hat{\pi}_j$
4. ΔD_j versus h_j
5. $\Delta\hat{\boldsymbol{\beta}}_j$ versus h_j

Plots number 4 to 5 will allow direct assessment of the contribution of leverage to the value of the diagnostic statistic.

Classification tables

A classification table often referred to as a confusion matrix, will be used to summarise the outcomes of the fitted logistic regression model. The outcome variable, and our dichotomous variable, Loan Default, whose values are obtained from the predicted logistic probabilities, will be cross-classified to produce this table. Instead of calculating the likelihood of the occurrence, the model's coefficients are employed in this application to forecast the outcome (in a binary sense).

Establishing a cut-off point and contrasting each estimated probability is necessary to create the dichotomous variable. The derived variable is set to

1 if the calculated probability exceeds 0.5; otherwise, it is set to 0. By adjusting the model's Specificity and Sensitivity, we will be able to determine its effects on the borrower and the bank. This discussion will be done using the confusion matrix.

In order to get the best classification model possible, one of our goals is to choose a cut-off point that maximises both sensitivity and specificity. A graph showing sensitivity and specificity versus various cut-off points will aid in decision-making.

Evaluation measures from the confusion matrix

As already discussed above, the confusion matrix is a 2×2 table that contains four outcomes produced by a binary classifier. These outcomes are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The true positive indicates a situation where the classifier correctly predicts a loan application as a defaulter, and the true negative occurs when the classifier correctly predicts a non-defaulter. Both of these outcomes are ideal for the bank. Conversely, a false positive means that the classifier wrongly predicts a non-defaulter as a defaulter. This outcome, though not ideal, is relatively less costly to the bank. The direct consequence to the bank is the rejection of an otherwise good borrower and the subsequent loss of business.

In some cases, this loss of business can be mitigated by the bank investing such available funds into non-risky assets such as Government Bonds and Treasury Bills and gaining better margins on interest. A False negative implies the classifier wrongly predicts a defaulter as a non-defaulter. The implications of this misclassification are dire for the bank. This means the bank will grant loans to bad customers. The direct consequence of this misclassification is that depositors' funds given to bad customers are not paid back. Approving loans to bad customers will pose serious problems such as liquidity risk, compliance issues, profitability challenges for the bank, and, in some cases, bank bankruptcy. This is the area where this research will highlight the importance of using the optimal cut-off point as a reference for loan approval decisions by the bank.

Various measures, such as "error rate", "accuracy", "specificity", "sensitivity", and "precision", are derived from the confusion matrix. These measures are usually expressed in percentages. Please see Table 6 below.

Table 6. Summary Diagnostic Testing Accuracy Measures

Measure	Formula
Percentage Correctly Classified (PCC) or Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Misclassification rate or Error rate (1 – Accuracy)	$\frac{FP+FN}{TP+TN+FP+FN}$
Sensitivity (or Recall)	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision (or Positive Predictive Value)	$\frac{TP}{TP+FP}$
Negative Predictive Value	$\frac{TN}{TN+FN}$

(Akosa, 2017)

Accuracy is the percentage of correct predictions our classifier has made on the testing dataset. The Error Rate is the percentage of misclassification (false predictions) that our classifier has made. Both measures range from 0% to 100%. An increasing or high Accuracy is ideal, while a decrease or low Error Rate is ideal.

Sensitivity (or Recall) is the ability of our classifier to correctly predict the number of loan defaulters from the total number of loan defaulters. Conversely, Specificity is the ability of our classifier to correctly predict the number of loan non-defaulters from the total number of loan non-defaulters. Both these measures also range from 0% to 100%. A high measure is ideal in both measures, however, in our case where we have an imbalanced data (92% loan non-defaulters against 8% loan defaulters), the measures are going to be imbalanced, hence specificity will be a lot higher than sensitivity.

Precision (Positive Predictive Value) is the percentage of predicted loan defaulters who are actual loan defaulters. This shows how many of the predicted loan defaulters are actual loan defaulters. Conversely, "Negative Predictive Value" is the percentage of predicted loan non-defaulters who are actual loan non-defaulters. This shows us how many predicted loan non-defaulters are actual loan non-defaulters. If both these measures are higher (as close to one hundred as possible), then it suggests that our classifier is good. In highly imbalanced data, this measure may be misleading.

The area under the receiver operating characteristic curve (ROC Curve)

Sensitivity, specificity, and other classification performance indicators derived from a 2×2 table are dependent on the single cut-off point used to classify a test result as positive. On the other hand, the chance of identifying real positive rates (sensitivity) and false positive rates (1-specificity) is plotted on the ROC Curve for various potential cut-off points. The ability of the model to distinguish between participants who experience the outcome of interest and those who do not is measured by the area under the ROC curve, which has a range of 0.5 to 1.0. You can find general talks about ROC curves in (Altman, 1991; Krzanowski & Hand, 2009; Swets, 1996 Zhou et al., 2002). (Gehlbach, 1988) gives an illustration of its application.

The following common rule of thumb describes the area under the ROC curve that indicates good discrimination (Hosmer et al., 2013).

$$\text{“if} \begin{cases} ROC = 0.5 & \text{This suggests no discrimination, so we might as well toss a coin.} \\ 0.5 < ROC < 0.7 & \text{We consider this poor discrimination.} \\ 0.7 \leq ROC < 0.8 & \text{We consider this acceptable discrimination.} \\ 0.8 \leq ROC < 0.9 & \text{We consider this excellent discrimination.} \\ ROC \geq 0.9 & \text{We consider this outstanding discrimination.} \end{cases}$$

”

Chapter Summary

The introduction chapter provides an overview of the study's data characterization and the methods used to construct proposed models for analyzing the data. The study focuses on borrowers with loans from a commercial bank in Ghana, and the data is obtained from one of the indigenous commercial banks. The data consists of 9,590 borrowing customers, with ten variables for each borrower, including loan details, borrower information, and loan default status.

The study aims to develop 16 binary models using various algorithms, including Bayesian Network, Decision Tree, Support Vector Machine, Random Forest, and others. These models will be grouped into five categories based on their underlying methodologies.

To build and validate the models, the data set will be divided into a training sample (70%) and a test sample (30%). The logistic regression model and the other 16 models will be developed using the training sample and applied to the test sample to compare their performance using the area under the curve (AUC).

Data characterization includes categorizing the dependent variable "Loan Default" and nine independent variables, including gender, marital status, employment type, loan description, economic sector, amount financed, loan tenor, interest rate, and borrower's age. The study uses logistic regression due to its suitability for binary dependent variables and categorical independent variables, making it the ideal choice for the analysis.

Overall, the introduction chapter sets the groundwork for the study, explaining the data used, the methods applied, and the rationale behind the chosen statistical techniques to analyze and model credit risk in Ghana's commercial banks.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

In this chapter, we present and discuss the results of the findings of the study. The first section of this chapter presents the summary of the descriptive statistics of the variables (Categorical and Numerical) as well as the graphical descriptive analysis of the borrowers. The chapter goes on to explain how the data on borrowers is modelled using the logistic regression model, taking into account the characteristics of the data under consideration. The subsequent sections present results including the findings from the model employed, the diagnostic analyses, which were completed by employing the Classification Table, ROC analysis and evaluation measures of the classification table to verify the model's robustness.

Summary of Descriptive Statistics of the Variables

Table 23 to 26 at appendix presents the descriptive statistics of the categorical variables. Table 23 shows that out of the total of 8,687 borrowers, 652 borrowers defaulted on loans granted to them. The loan defaulters represent only 7.5% of the total number of borrowers. The condition in which the class of loan defaulters is under-represented in relation to the class of loan non-defaulters is a fairly prevalent situation in real-life credit rating systems (García & Sánchez, 2013) which is the case with the observations in this study.

Observation from the gender statistics inTable 24 indicates that the number of males (5,406) is 24.4% higher than the number of females (3,281). According to (Amu, 2005), women lack access to credit because of their relative low level of savings, low income, nature of business engaged in, collateral etc. This argument is not different from that of the Berton Woods' assertion on debt and gender equality which states that "women face multiple and intersecting forms of discrimination based on age, race, income levels, ability and disability, gender identity and sexual orientation".

Similarly, marital status is categorised into these segments; divorced, engaged, married, single or widowed. It can be seen from Table 25 that borrowers who were married or single form more than 90% of the total number of borrowers considered in the study. However, borrowers who were divorced or engaged formed less than 10% of borrowers with widowed borrowers forming the least category.

Also, the type of employment of the borrower is categorised into the following: full time, part time, retired or other. The "other" category consists of contractors, freelancers, interns, consultants, and self-employed individuals. Statistical findings from Table 26 shows that full time employees form 96.4% of borrowers considered in the study. Conversely, the other categories form the remaining 3.6%. Commercial banks are comfortable lending to full time employees than any category of employee; this is because the source of income for repayment can be easily verified.

By putting the loan description in perspective, the commercial bank under consideration categorises its retail loan portfolio into five categories, namely, CAGD Salary Loan Scheme, Commercial Loans, Institutional Workers Loan Scheme, Micro Credit Loan Scheme or Salaried Workers Loan Scheme. From Table 27, it can be seen that the borrowers who have their source of repayment from their salaries form 93.9% of the total number of borrowers. Salaried workers are prevalent because their source of income for repayment is easily verifiable, and the deductions are mainly from source. For example, employees on government payroll in Ghana are paid through the CAGD and hence, the banks have first-hand access to funds before the remainder are paid as salaries.

In addition, the borrowers are segmented into the three sections of the economy, namely, primary, secondary, and tertiary. These can also be classified as the agricultural, manufacturing and service sectors. From Table 28, it can be deduced that the borrowers are from the tertiary sector (service) of the economy which comprises of media, education, banking, tourism, insurance, transportation, etc. represent 98.8% of the total number of borrowers. The remainder of the borrowers in the primary and secondary sectors of the economy form 1.2%.

Table 7 below displays the summary statistics for the continuous variables. From the table mentioned, the mean age of borrowers is approximately 40 years and the mean loan amount received by the borrowers is about GH¢13,236 at a mean interest rate of about 28% p. a. for a mean tenor of about 3 years. A skewness of 6.10 and a kurtosis of 75.5 reported on the amount financed variable implies that the distribution of the amount financed is highly positively skewed with a long right tail. The values of skewness and kurtosis of the amount financed variable signifies the presence of extreme values. We did not treat these extreme values as outliers because they represent vital information.

As a rule of thumb from (Bulmer, 1979) who stated that if skewness is between -0.5 and +0.5, the distribution is approximately symmetric. The reported skewness of Tenor of loan and Age of borrowers are -0.24 and 0.27 respectively, hence, the data on Tenor and Age are approximately symmetrical. Conversely, the interest rate charged by the bank shows some skewness, 1.5 and a kurtosis of 5.6 which implies a positive skewed data with long tails at both ends of the distribution. They also signify the presences of extreme values at both ends of the distribution. A detail analysis of the mean, median, mode, minimum value, maximum value, range, kurtosis, and skewness for each of the continuous variables including their histogram is displayed in the appendix.

Table 7. Descriptive Statistics of Numeric Variables

	Amount Financed (GH¢)	Tenor (Days)	Interest Rate (% p. a.)	Age (Years)
N	8,687	8,687	8,687	8,687
Mean	13,235.73	1,136.89	27.48	39.88
Median	8,217.60	1,094	28	39
Mode	5,127	1,094	28	37
Skewness	6.10	-0.24	1.45	0.27
Kurtosis	75.50	-0.10	5.55	-0.71
Range	469,624.85	2,524	26	63
Minimum	550.59	30	20	20
Maximum	470,175.44	2,554	46	83

Author’s Construction

Applying the Binary Logistic Regression Model

As previously stated, binary logistic regression offers several benefits over most other discriminating techniques. The relationship between the dependent and independent variables need not be linear, to begin with. Logistic regression can handle any relationship since it applies a non-linear log transformation to the projected odds ratio. Second, whereas multivariate normal independent variables offer a more stable solution, they are not required. Additionally, multivariate normally distributed error terms (also known as residuals) are not required. Finally, homoscedasticity is not required. Heteroscedastic variances are not required for each independent variable level in logistic regression. It can also handle nominal and ordinal data as independent variables. The independent variables do not have to be metric (interval or ratio scaled).

Despite the foregoing, "binary logistic regression" requires the fulfilment of certain assumptions. These assumptions are similar to multivariate regression, with the primary distinction being that binary logistic regression requires a binary dependant variable. The dependent variable in our study, loan default, meets this criterion. Second, because logistic regression implies that $P(Y = 1)$it represents the likelihood of an event occurring, the dependent variable must be appropriately coded. For instance, in a binary regression, the dependent variable's factor level 1 should indicate the intended outcome. The intended outcome of this research is for the borrower to default on the loan granted by the bank (this represents the presence of the condition under study, which is loan default). Thirdly, the model should be well fitted. There should be no over- or under-fitting. That is, not only should useful variables be included, but also all meaningful variables. The logistic regression is estimated using the forward stepwise likelihood ratio approach to meet this third requirement. Fourthly, the error terms must be independent. Each observation must be independent in order for logistic regression to work.

Additionally, the model should have little or no multicollinearity. That is, the independent variables should not be correlated. The absence of collinearity is identified in our study using two primary indices: the variance inflation factor (VIF) and spearman's correlation. The collinearity levels in our data are seen in Table 8 below using the VIF. The table exhibits a VIF range of 1.017 to 1.506 and a Tolerance range of 0.664 to 0.983. (A Tolerance measure of less than 0.10 signifies collinearity).

Table 8: Collinearity Statistics

Variable	Tolerance	VIF
Amount Financed	0.84	1.19
Tenor	0.66	1.51
Interest Rate	0.70	1.44
Age	0.99	1.02
Gender	0.75	1.34
Marital Status	0.97	1.03
Type Of Employment	0.83	1.20
Loan Description	0.75	1.36
Economic Sector	0.96	1.04

Author's Construction

As a general rule of thumb, a "VIF of 1 signifies no correlation, a VIF of between 1 and 5 signifies moderate correlation and a VIF greater than 5 signifies high correlation" (Rawlings et al., 1998; Weisberg, 2005; Yadolah, 2008). Little collinearity between the independent variables is seen in Table 6 above. The Spearman's correlation analysis table, Table 23, which is included in the appendix, also supports this. Fifth, logistic regression assumes log probabilities and the linearity of independent variables. It demands that the independent variables have linear relationships with the log odds, even though it does not require linear relationships between the independent and dependent variables. Performing the Box-Tidwell test satisfies the sixth requirement (Box & Tidwell, 1962). The appendix's Table 24 demonstrates that the assumption of the logit's linearity has been satisfied. Finally, substantial sample size is needed. Maximum likelihood estimates require at least 10 cases for each independent variable in the analysis, whereas ordinary least squares only require five cases (e.g., simple linear regression, multiple linear regression). This is because maximum likelihood estimates are less accurate than ordinary least squares estimates. This assumption has also been met, as shown by looking at the frequency tables of the variables in the appendix.

Because the dependent variable is dichotomous and follows the Bernoulli distribution, a random selection using the random sampler in SPSS was used to select approximately 70% of cases for training the model and 30% for testing the model. Below shows the Bernoulli random segregation of the data into 70% train and 30% test cases for the formulation and validation of the model, respectively.

Table 9. Case Processing Summary

Cases	N	Percent
Selected Cases (Included in Analysis)	6,108	70.30
Unselected Cases (Test)	2,579	29.70
Total	8,687	100.00

Author’s Construction

The binary logistic model

Table 10 below shows the resulting model that was created using the data. Only five independent variables are statistically significant, according to the model. Tenor, Interest Rate, Age, Gender, and Loan Description are important independent factors. It should be mentioned that the Wald chi-square test was used to do statistically significant checks to see if the coefficients are equal to zero. The coefficients are all substantially different from zero at the significant threshold of 0.05, as observed.

Table 10. Variables in the Equation

	B	S.E.	Wald	df	Sig	Exp (B)
Tenor	-0.001	.000	35.820	1	0.000	0.999
Interest Rate	0.214	.027	63.642	1	0.000	1.238
Age	-0.011	.005	4.701	1	0.030	0.989
Gender (1)	-0.633	.149	17.908	1	0.000	0.531
Loan Description (2)	-1.213	.284	18.294	1	0.000	0.297
Loan Description (3)	-1.628	.174	87.815	1	0.000	0.196
Loan Description (4)	-3.782	.407	86.128	1	0.000	0.023
Constant	-5.289	.827	40.894	1	0.000	0.005

Author's Construction

The coefficients for the (fitted) line and other pertinent details about the coefficients are contained in the Variables section of the equation table.

The line obtained from the output has the following equation:

$$g(x) = -5.289 - 0.001x_1 + 0.214x_2 - 0.011x_3 - 0.633x_4 - 1.213D_{52} - 1.628D_{53} - 3.782D_{54} \quad (20)$$

where x_1 = Tenor, x_2 = Interest Rate, x_3 = Age, x_4 = Gender and D_{52}, D_{53}, and D_{54} are the categories of Commercial loan, Institutional Workers Loan or Micro-credit Loan schemes respectively in the Loan Description variable.

The column labelled "B" in the table above represents the coefficients of the respective variables. The interpretation of the negative coefficients associated with the continuous variables Tenor and Age is that the higher the Tenor or Age, the less likely a borrower will default on a loan granted. Conversely, the interpretation of the positive coefficient of Interest Rate is that the higher the interest rate, the more likely a borrower will default. The negative coefficients associated with the categorical variables, Gender (1) [Gender (1) here denotes the first category in the grouping, which is female.], for example, means that a female borrower is less likely to default on a loan. Borrowers in the second, third or fourth category of the variable Loan Description are all less likely to default on loans granted to them.

A further review of Table 10 also indicates other measures like the odds ratio represented by Exp(B) in the last column of the table. The interpretation of the odds ratio for the variable Tenor is that, for every unit increase in the tenor of the loan, the odds of a borrower defaulting on a loan is decreased by 1%. The odds ratio of 1.238 for the variable Interest Rate means that for every unit increase in an interest rate of a loan, the odds of a borrower defaulting on a loan is increased by 23.8%. In the case of the variable, Age, the odds of a borrower defaulting on a loan decrease by 1% at every unit increase in the age of the borrower. The output further reveals that female borrowers are 53% less likely to default on loans than male borrowers; conversely, a male borrower is 1.89 times more likely to default than a female borrower. In other words, the probability of default for a female borrower is 0.35, and the probability of default for a male is 0.65. Finally, if a borrower belongs to any of the loan schemes, Commercial, Institutional Workers or Micro-Credit, they are less likely to default on a loan by 29.7%, 19.6% or 2.3%, respectively.

Goodness-of-fit Statistics

The SPSS output displays the "Omnibus Tests of Model Coefficients" and the Hosmer and Lemeshow test. The Omnibus Tests of Model Coefficients (See Table 11 below) are used to test the model fit. This is the overall Chi-square test of fitness of the model under the hypothesis:

$$H_0: \beta_j = 0 \text{ for all } j$$

$$H_1: \beta_j \neq 0 \text{ for at least one coeffcient}$$

Table 11. Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step	4.738	1	.030
Block	1,287.061	9	.000
Model	1,287.061	9	.000

Author's Construction

Since there is a considerable improvement in fit relative to the null model in our investigation, the model is significant and therefore demonstrates a strong match. The overall model is statistically significant $\chi^2(9) = 1{,}287.06, p < 0.05$

The "Hosmer-Lemeshow test" examines the null hypothesis that the model's predictions match the observed group memberships. A chi-square statistic is calculated when comparing the observed frequencies to those predicted by the linear model. A no significant chi-square means that the data were well fitted to the model.

Table 12. Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
6	4.096	8	.848

Author's Construction

Here, the model adequately fits the data. Hence, there is no difference between the observed and predicted models.

Model Summary

The Model Summary shows the Pseudo R-Square. Pseudo means that it is not technically explaining the variation. But they can be used as approximate variation in the criterion variable.

Normally used is Nagelkerke's *R-Square*. This is an adjusted version of the Cox & Snell *R–square* that adjusts the scale of the statistic to cover the full range from 0 to 1.

Table 13. Model Summary

-2 Log likelihood	Cox & Snell R-Square	Nagelkerke R-Square
1,921.171	0.190	0.465

Author's Construction

In this case, from

Table 13 above, it can be said that a 46.5% change in the dependent variable can be accounted for by the independent variables in the model.

Model Diagnostics

The model fits, as evidenced by the summary statistics above. As a result, we do not anticipate that a diagnostics analysis would uncover several poorly fitting cases. However, we may identify a few cases that do not fit or have a significant effect on the estimated parameters. The key plots are given in Figure 1 to Figure 5 below. We discuss each plot in turn.

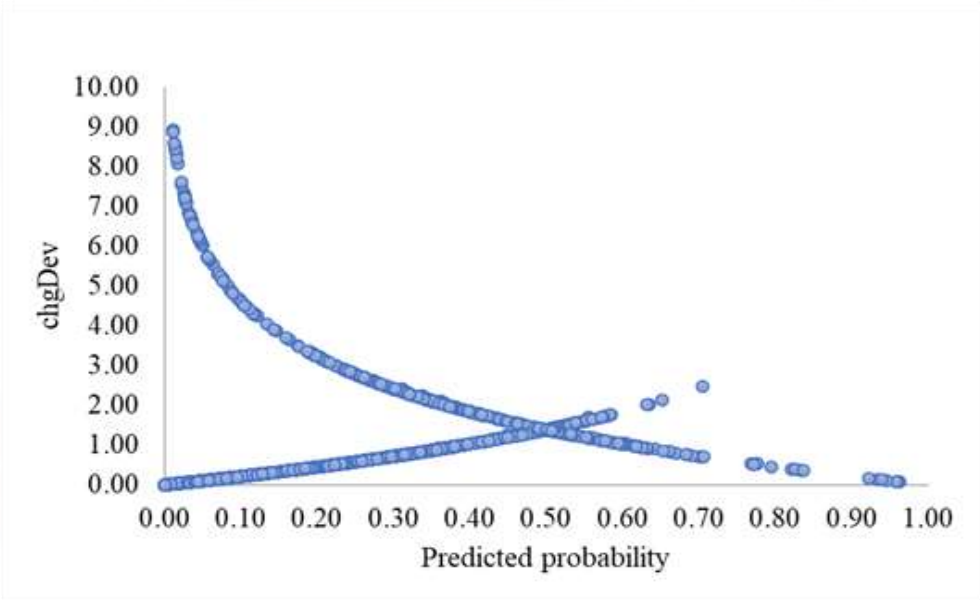

Figure 1. Plot of change in deviation versus the predicted probability from the fitted model

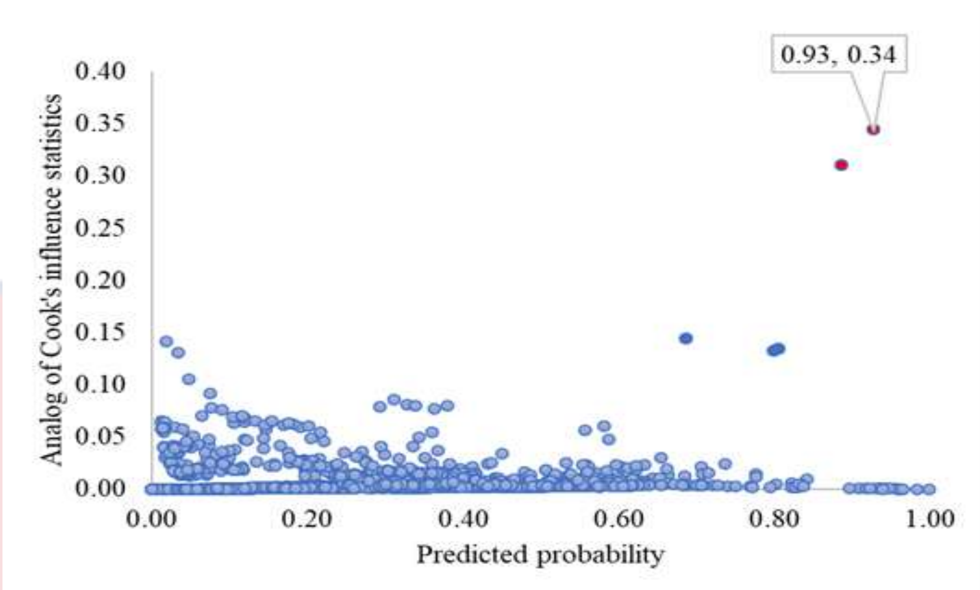

Figure 2. Plot of Cook's distance versus the predicted probability from the fitted model

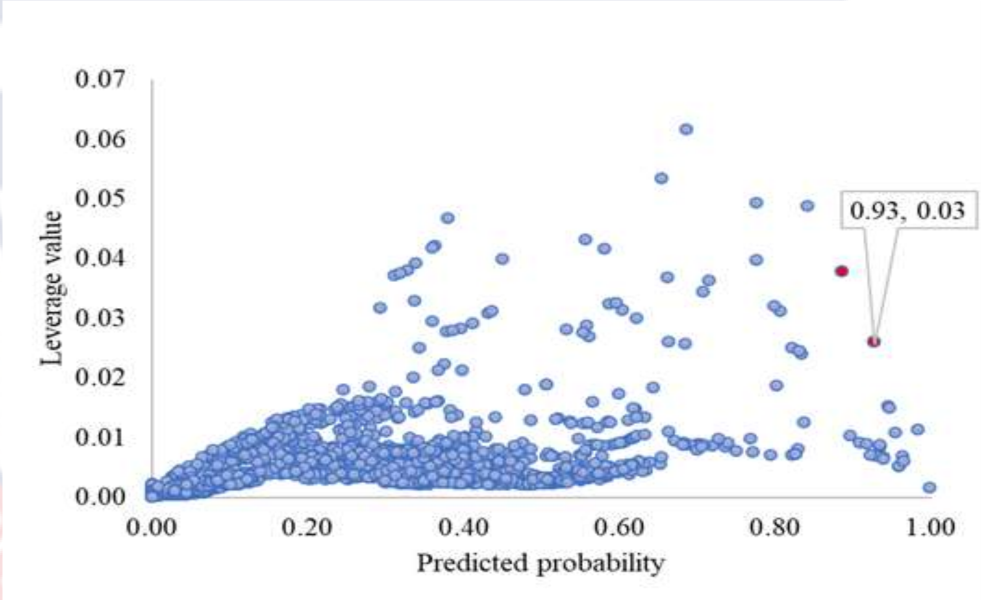

Figure 3. Plot of leverage versus the predicted probability from the fitted model

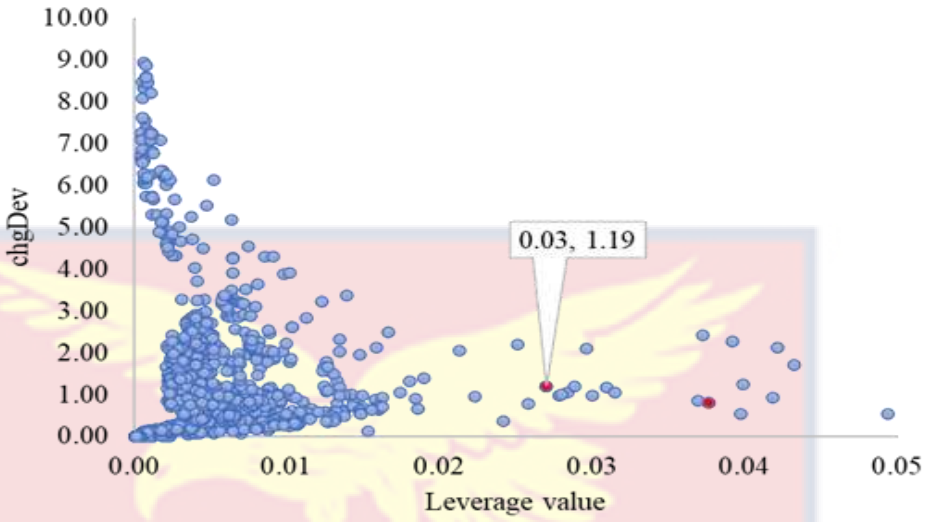

Figure 4. Plot of change in deviation versus the leverage values from the fitted model

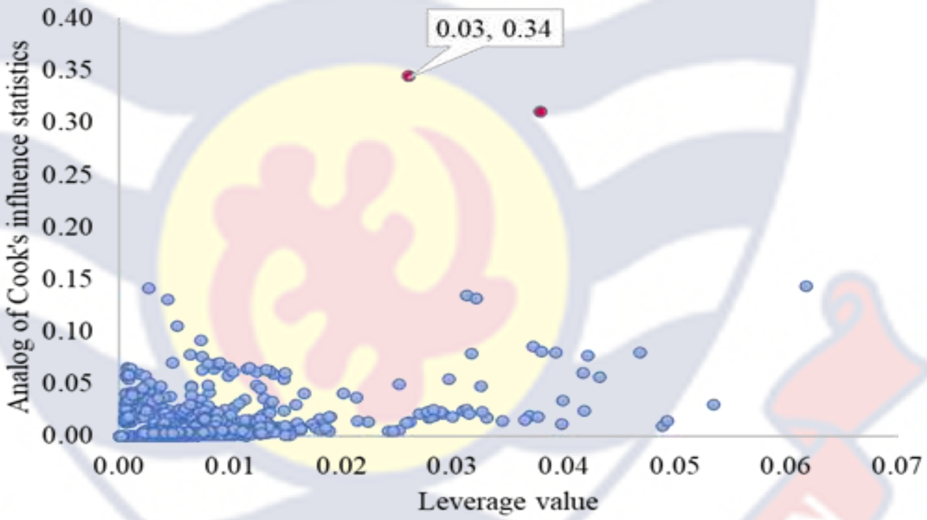

Figure 5. Plot of Cook's distance versus the leverage values from the fitted model.

Figure 1's plot form exhibits quadratic-like curves. We may find scenarios when the model does not adequately explain the data using the change in deviation plot. Poorer fits are indicated by greater fluctuations in deviation. A curve that runs from the lower left to the higher right and a curve that runs from the top left to the lower right are the two main patterns in the plot. Points lying in the top left or top right corners of plots are frequently used to depict cases that are poorly fitted in general. When the dependent variable has a value of 0, the curve that runs from the bottom left to the top right represents such situations. As a result, the model does not adequately account for loan non-defaulters with high model-predicted odds of default. Figure 1 shows that only a small number of cases had probability predicted by the model that were higher than 0.6 and matching changes in deviance values that were less than 4. This shows how well the model predicts loan non-defaulters. On the other hand, the curve that runs from upper left to lower right is what happens when the dependent variable is equal to 1. As a result, the model does not adequately account for loan defaulters with low model-predicted odds of default. The change in deviation for cases with model-predicted probabilities of less than 0.1 similarly rises dramatically toward nine, as shown in Figure 1. These instances show loan defaulters whom the model did poorly fitting. (Hosmer et al., 2013) utilise 4 as a rough approximation to the upper 95% of the change in deviation distribution because, under m-asymptotics, these numbers would be distributed roughly as $\chi^2(1)$ with $\chi^2_{0.95}(1) = 3.84$ and conclude that if most values of change in deviance are less than 4 then the plots show that the model fits reasonably well. A further inspection of Figure 1 shows that the majority of values of change in deviance are less than 4, hence, the plots show that the model fits reasonably well.

The influence diagnostic in Figure 2 shows a plot of projected probability versus Cook's distance. With a few small exceptions, the Cook's distances plot often has the same form as Figure 3. These exceptions have a lot of leverage and can affect the analysis. From the data as a whole, two (2) situations stand out (highlighted in Figure 2). Since each number is less than

0.4, they are not exceptionally large in themselves. In order for a particular example to have an impact on the calculated coefficients, the influence diagnostic needs to be greater than 1.0, according to (Hosmer et al., 2013; Pituch & Stevens, 2015), among others. Nevertheless, there are always exceptions, and regardless of the precise magnitude, it is best practise to note Cook's distance values that are outliers. The example with the highest influence diagnostic in Figure 2 (Data callout), 0.34, is not the one with the highest value of change in deviation, as indicated in Figure 4. This case has a mild lack of fit and moderate leverage (Figure 4). The leverage values at the other places in this similar region of projected probability are lower, but the change in deviance values are only somewhat higher.

The contribution of leverage to the value of the change in deviance and influence can be directly assessed using Figures 4 and 5. Both of the aforementioned figures demonstrate that the majority of cases with higher leverage often exhibit a moderately low change in deviation as well as a relatively low impact on the overall model. Despite the aforementioned, Figure 4 shows that many examples with low leverage values have very substantial change in deviation, which indicates a lack of fit for these circumstances.

These diagnostics can identify which instances have an impact on each component of the model, as was covered in chapter three. There shouldn't be any lack of fit brought on by incorrect transformations after properly modelling predictor transformations. It is frequently the case, though, and our financial data is instructive, that extreme predictor values can nevertheless have an impact on estimations of coefficients incorporating the predictor. It is irrational to believe that given real data all points will agree with the overarching model because these ostensibly relevant situations are merely random occurrences. This discovery supports our choice not to treat extreme values as outliers that should be removed from the data because they have an impact and leverage on the overall model's prediction ability.

Classification table

The output result's classification table summarises the classification of the observed groups and the expected groups. See Table 14 below.

Table 14. Classification Table (Default cut-off point of 0.5)

Observed		Predicted					
		Selected Cases			Unselected Cases		
		DEFAULT		Percentage Correct	DEFAULT		Percentage Correct
		NO	YES		NO	YES	
LOAN	NO	5,594	**65**	98.9	2,342	34	98.6
DEFAULT	YES	**347**	102	22.7	157	46	22.7
Overall Percentage				93.3			92.6

Author's Construction

Our Logistic regression model estimates the probability of default occurring at an even chance of occurrence or cut off point of 0.5. If the estimated probability of default occurring is greater than or equal to 0.5, the model classifies the borrower as a loan defaulting borrower. Conversely, if the probability of default is less than 0.5, then the model classifies the borrower as non-defaulting.

It can be seen from the table that for the selected cases (train data), the specificity which is the True Negative Rate is 98.9% while the sensitivity, the True Positive Rate is 22.7% with a stated overall model performance of 93.3%. This trend in classification is extended to the unselected (test data) cases with specificity of 98.6% and sensitivity of 22.7% with a slightly reduced overall performance percentage of 92.6%. There is clearly an imbalance in classification using the even chance of occurrence or cut-off point of 0.5.

The interpretation for false-negative in this case is that the model has classified 347 borrowers as non- loan defaulters when in actual fact they defaulted. The consequence of this misclassification is that funds given to these 347 borrowers are potentially lost, irrecoverable and may be written off affecting profitability of the bank and the aggregate consequence being that the bank becomes bankrupt.

On the contrary, the interpretation of the false-positive is that sixty-five borrowers were misclassified as loan defaulters when in actuality these borrowers did not default. The consequences of false-positive for the bank are that these sixty-five are denied loans which may lead to the loss of these customers and business (opportunity cost of not giving these good borrowers loans that will augment the interest income of the bank). Some of this opportunity cost may be recovered by the bank when the bank decides to invest such available funds in non-risky investments such as treasury bills and Government bonds, in some cases, these risk-free bills and bonds yield a better margin than lending to customers.

The above-described circumstances present a practical challenge to the bank whose main aim is to reduce credit risk. To overcome this problem of imbalance in classification, we propose that a plot of sensitivity and specificity versus predicted probability from the model is constructed. The two curves obtained from this plot intersect and this intersection point should be considered the optimal cut-off point for the loan decision making of the bank. See Figure 6 below.

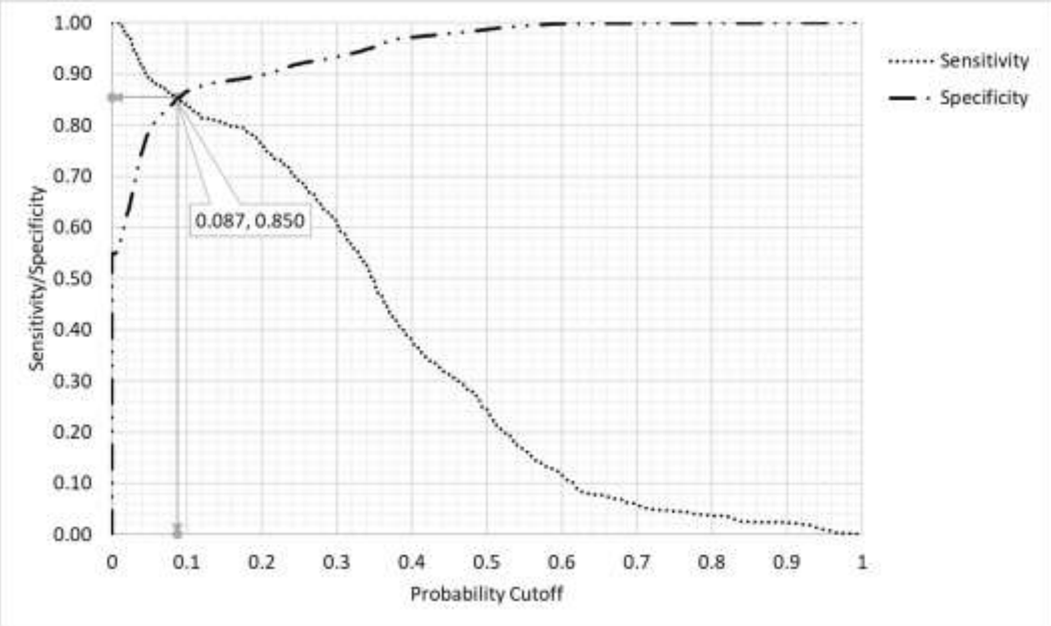

Figure 6. Sensitivity-Specificity vs Predicted Probability

This “optimum cut-off point” of 0.087 deduced from the plot reduces the number of false-negatives predictions of the model thereby reducing credit risk as shown in the classification table below.

Table 15. Classification Table (Optimum cut-off point of 0.087)

Observed		Predicted					
		Selected Cases			Unselected Cases		
		DEFAULT		Percentage Correct	DEFAULT		Percentage Correct
		NO	YES		NO	YES	
LOAN DEFAULT	NO	4,831	**828**	85.4	2,008	368	84.5
	YES	**65**	384	85.5	35	168	82.8
Overall Percentage				85.4			84.4

Author’s Construction

This balances the predictive power of the model with specificity and sensitivity both reporting a classification accuracy of about 85% for the selected cases. Credit risk is clearly reduced since the number of false negatives has reduced from 347 to 65, improving sensitivity from 22.7% to 85.5%. A similar improvement is also seen in the test cases.

ROC curve analysis

In other words, it gives a measure of the model's capacity to distinguish between borrowers who will fail on their loans and borrowers who won't in our scenario. An ROC curve is a measure of goodness-of-fit frequently used to assess the fit of a logistic regression model. As a result, cases can be classified using the generated logistic regression model and the ROC curve. The best cut-off point for the logistic regression model can also be found using the ROC curve. The ROC curve is created by plotting sensitivity versus 1-specificity, as Chapter 3 of this paper explains. The output of the ROC curve plotted is presented below.

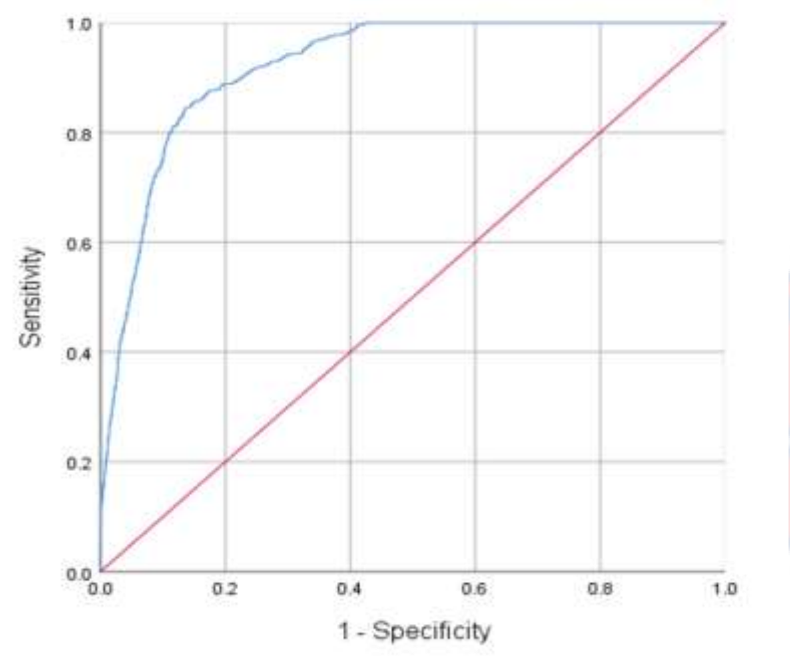

Figure 7. ROC Curve

A model with no predictive value will have a slope of one, resulting in an ROC value of 0.5. A close inspection of the curve in Figure 7 above will reveal that the part of the curve nearest to the sensitivity value of 1 is about 0.85 with corresponding inverse of specificity (1-specificity) of about 0.15. This point on the curve corresponds to the optimal cut-off point of 0.087

obtained in Figure 6 above. A table of all the possible cut-off values and their corresponding sensitivity and inverse of specificity values is presented at the Appendix, an inspection of which will yield the optimal cut-off point.

The Area Under the ROC curve which measures the accuracy of our model is presented below in Table 16 and Figure 8.

Table 16. Area Under the Curve of the ROC

			Asymptotic 95% Confidence Interval	
Area	Std. Error	Asymptotic Sig.	Lower Bound	Upper Bound
0.921	0.004	0.000	0.914	0.929

Author's Construction

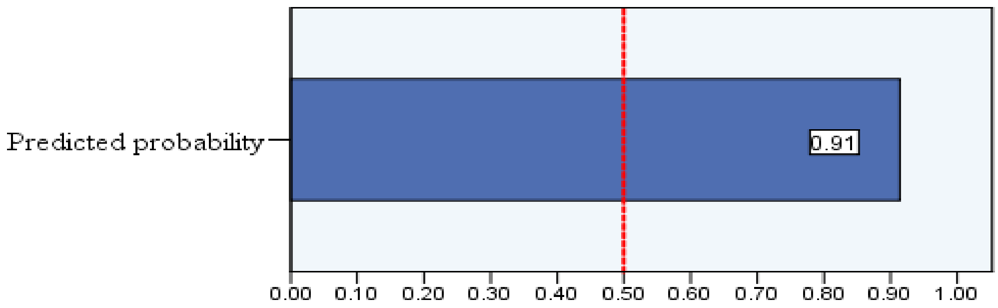

Figure 8. Estimated Accuracy of the Model

It can be seen that the asymptotic significance of the model is less than 0.05, hence, the model is a good classifier. The model is a good classifier/discriminator when applying the general principle from (Hosmer et al., 2013), which is covered in Chapter three of this study.

Effect of the Optimal Cut-off Point on the Logistic Regression Model

The table below gives us a summary and an indication of the effect of the "optimal cut-off point" on some accuracy measures, including our model's sensitivity and specificity.

Table 17. Summary of Accuracy Measures

Measure	Cut-off 0.5	Cut-off 0.087
Percentage Correctly Classified (PCC) or Accuracy	93.3%	85.4%
Misclassification rate or Error rate (1 – Accuracy)	6.7%	14.6%
Sensitivity (or Recall)	22.7%	85.5%
Specificity	98.9%	85.4%
Precision (or Positive Predictive Value)	61.1%	31.7%
Negative Predictive Value	94.2%	98.7%

Author's Construction

The other accuracy metrics all appear to have decreased, with the exception of sensitivity and negative predictive values, which improved using the optimal cut-off point. This statement is consistent with discussions made by (Akosa, 2017) that these measures may be deceptive if the unbalanced nature of the data on the dependent variable is not taken into consideration. Only 7.5% of a total of 8,687 borrowers actually defaulted on loans, according to statistics acquired from the bank. The model's capacity to distinguish between true positives and true negatives was significantly impacted by this mismatch. Despite the fact that

Table **17**'s error rate increased by two-and-a-half times, from 6.7% to 14.6%, the classification of false positives and the enormous number of loan non-defaulters are the key factors in play. The precision value, which appears to have been cut in half, is supported by the same argument.

The effect of shifting the cut-off points on the importance of the independent variables in model were investigated. A selection of a cut-off point does not affect the influence or importance of the independent variables on the dependent variable, it merely establishes a point where a borrower may be rejected or approved. Hence the choice of cut-off point affects the number of borrowers that the bank will grant loans. See Figure 9 below.

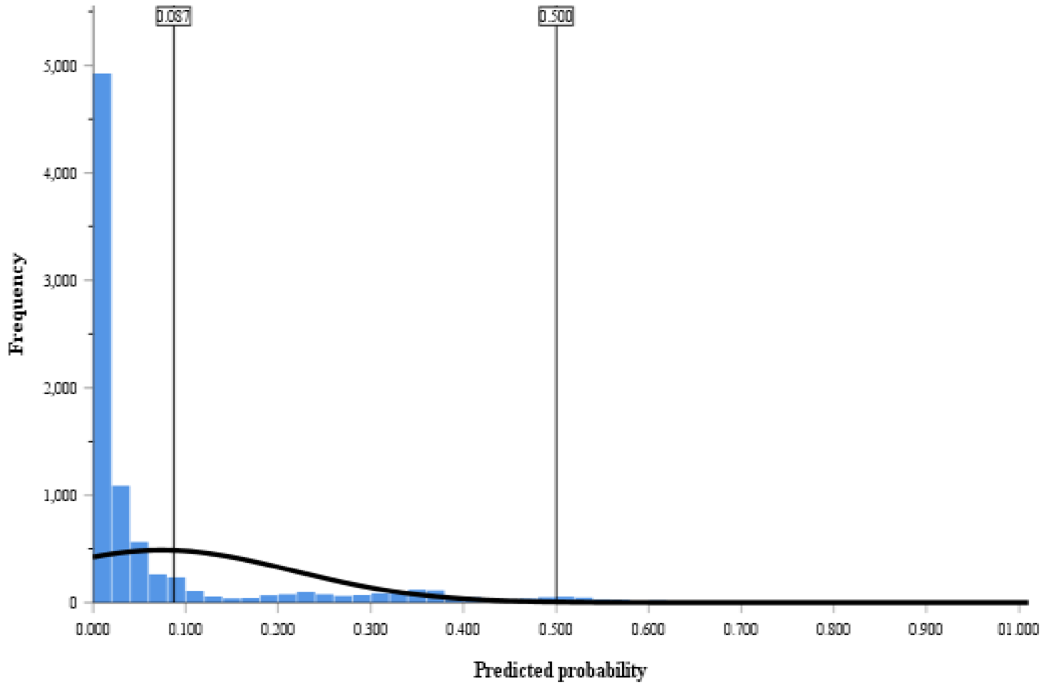

Figure 9. Histogram of Predicted Probability

Figure 9 shows that at a cut-off point of 0.5, only a small number of borrowers—247 out of 8,687 borrowers—are labelled as loan defaulters. Using the optimal cut-off value of 0.087, the number of loan defaulters rises to 1,748 (FP+TP). It is anticipated that the credit portfolio quality, or the NPL ratio, will improve since the optimal cut-off point appropriately categorises more loan defaulters. Additionally, it can be observed that even with a cut-off point of 0.087, about 80% of borrowers still receive loans with credit risk that is at its lowest possible level. In a case where the bank intends to reduce credit risk to the bearest minimum and maintain a respectable NPL ratio, the upper limit for granting a loan should be the optimal cut-off point determined. However, if the bank considers increasing its customers' access to credit, then the bank can set the optimal cut-off point as the lower limit for approving loans and the default cut-off point of 0.5 as the upper limit. A threshold the bank can operate and expect a reasonably good reduction in credit risk beyond which the bank might as well grant every borrower a loan without considering the score as is currently the practice.

Comparison of the Logistic Regression Model with other Binary Models

The performance of each model, measured by their respective AUCs are summarized in the table below with the best performing alogorithm in each group is highlighted.

Table 18: Performance of Selected Models

	No Resampling	Oversampling	Undersampling	Rank
Models	AUC	AUC	AUC	Average AUC
Random Forest	**0.934**	**0.997**	**0.944**	**0.958**
Tree-AS	0.941	0.955	0.931	0.942
CHAID	0.941	0.950	0.927	0.939
C5	0.886	0.983	0.919	0.929
C&R Tree	0.864	0.908	0.905	0.892
XGBoost Tree	0.940	0.687	0.682	0.770
Random Trees	0.609	0.482	0.623	0.571
Logistic Regression	**0.921**	**0.920**	**0.929**	**0.923**
LSVM	0.921	0.919	0.914	0.918
Discriminant	0.755	0.749	0.767	0.757
XGBoost Linear	0.915	0.662	0.649	0.742
Neural Net	**0.927**	**0.942**	**0.942**	**0.937**
SVM	0.918	0.942	0.938	0.933
Bayesian Network	0.923	0.935	0.930	0.929
QUEST	**0.842**	**0.832**	**0.893**	**0.856**
Decision List	0.891	0.754	0.831	0.825
KNN	**0.879**	**0.978**	**0.915**	**0.924**

Author's Construction

Based on the AUC measure alone, and with reference to the common rule of thumb that indicates good discrimination by (Hosmer et al., 2013), the logistic regression model appears to perform almost at par with the top performing binary models. A clearer picture of the performance of the models is shown when the specificity and sensitivity of the models under each treatment are shown as highlighted in Table 19 to Table 21 below:

Table 19: Performance of Selected Models with No Resampling Applied

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
Random Forest	95.093	98.520	52.792	0.934
Logistic Regression	**93.337**	**98.600**	**22.700**	**0.921**
Neural Net	92.887	97.533	35.533	0.927
QUEST	93.572	99.877	15.736	0.842
KNN Algorithm	94.256	98.766	38.579	0.879

Author's Construction

Table 20: Performance of Selected Models with Oversampling Applied

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
Random Forest	99.398	98.784	100	0.997
Logistic Regression	**87.375**	**85.312**	**89.639**	**0.920**
Neural Net	87.958	85.471	90.526	0.942
QUEST	77.995	88.457	67.515	0.832
KNN Algorithm	96.298	94.454	100	0.978

Author's Construction

Table 21: Performance of Selected Models with Undersampling Applied

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
Random Forest	87.469	85.990	89.063	0.944
Logistic Regression	**86.967**	**86.154**	**89.691**	**0.929**
Neural Net	87.970	87.766	90.777	0.942
QUEST	86.967	87.432	87.379	0.893
KNN Algorithm	87.719	84.817	93.137	0.915

Author's Construction

It can be seen from Table 19 that the models' sensitivity values are much lower than their respective specificity values. This is indicative of models when they are applied to imbalanced data.

Table 20 and Table 21 show that applying resampling techniques improves the sensitivity values of the models, however, there is a slight dip in the accuracy measures of the models.

We suspect that resampling techniques, which involves the manipulation of the data in one way or another, introduces some bias into the data in that the resulting data set is not indicative of the natural structure of the data.

Table 22 below highlights the performance of the logistic regression model at the derived optimal cut-off point of 0.087

Table 22: Performance of Logistic Regression at the Optimal Cut-off Point of 0.087

Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
84.4	84.5	82.8	0.921

Author's Construction

Although the performance of the logistic regression model as show in Table 22 is slightly lower than the performance when resampling is applied, we do not believe that the difference in in performance is statistically significant.

In contrast to the other models employed in this study, it is worthy to note that logistic regression performs similarly even without applying resampling techniques. Additionally, logistic regression holds distinct advantages over the others; It enables the identification of significant borrower characteristics, simplifies the interpretation of odds and probabilities, and provides a relatively straightforward threshold for decision-making purposes.

Chapter Summary

In this chapter, the author presents and discusses the results of the study, focusing on the findings obtained from the logistic regression model applied to the data on borrowers. The chapter begins by providing a summary of the descriptive statistics for both categorical and numerical variables related to borrowers. These statistics include information about loan defaulters, gender distribution, marital status, type of employment, loan description, and the sectors of the economy borrowers belong to.

The authors then proceed to discuss the application of the binary logistic regression model to the data, explaining its advantages and assumptions. They highlight the importance of meeting these assumptions to ensure the model's accuracy. The model includes five significant independent variables: Tenor (loan duration), Interest Rate, Age, Gender, and Loan Description. The coefficients and odds ratios of these variables are analyzed to understand their impact on loan default probability.

To evaluate the model's goodness-of-fit, the authors use various statistical measures, including the Omnibus Tests of Model Coefficients and the Hosmer-Lemeshow test. They find that the model is statistically significant and well-fitted to the data.

The chapter also covers model diagnostics, examining specific cases that might not fit well or have a significant effect on the model's estimated parameters. The authors illustrate these diagnostics through various plots, such as change in deviation plots and influence diagnostics.

Next, the authors explore the classification table, which shows the model's ability to predict borrowers who will default on their loans accurately.

They discuss the issue of imbalance in classification and propose an optimal cut-off point (0.087) to address this problem. The optimal cut-off point balances sensitivity and specificity, leading to a reduction in credit risk.

The chapter further evaluates the model's performance using ROC curve analysis, which measures the model's ability to distinguish between defaulting and non-defaulting borrowers. The area under the ROC curve (AUC) is used as a measure of accuracy, and the logistic regression model is found to perform well compared to other binary models.

Finally, the authors compare the logistic regression model's performance with other models, including those that use resampling techniques to address the issue of data imbalance. While the resampling techniques improve sensitivity, the logistic regression model maintains competitive performance without applying such techniques. The authors emphasize the logistic regression model's advantages in identifying significant borrower characteristics, interpreting odds and probabilities, and providing a straightforward threshold for decision-making.

Overall, the chapter provides a comprehensive analysis of the logistic regression model's application to the borrower data, offering insights into credit risk assessment and loan default prediction for the bank under consideration.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

Overview

The rapid expansion of the banking sector in Ghana in recent times saw the country amass thirty-four (34) banks comprising of seventeen (17) foreign-controlled banks and seventeen (17) local-controlled banks. The regulator, BOG, in 2016 sanctioned an asset quality review program which revealed weaknesses in the banking sector. After BOG's banking sector clean-up implemented between the period of 2017 and 2018, the number of banks in Ghana shrank to twenty-three (23) banks comprising of fourteen (14) foreign-controlled banks and nine (9) local-controlled banks. The BOG reported that credit risk was the main factor that diminished the defunct banks' asset quality. It is interesting to note that all nine (9) defunct banks were local-controlled banks.

This research seeks to find out whether the existing data collected on borrowing customers of banks are adequate to use in a predictive model to reduce credit risk. The specific objectives were to identify and examine the factors contributing to credit risk among individual borrowing customers, propose a binary logistic regression model that discriminates loan defaulters from loan non-defaulters of credit facilities and determine the specificity and sensitivity of the discriminating model. The research further determines the optimal cut-off point which improves the sensitivity and investigates the effect of change in the cut-off point on the discriminating model and the number of borrowing customers.

The data was obtained from a Ghanaian local-controlled commercial bank's main banking application platform. The loan default is the dependent variable, while the borrower's age, gender, employment type, and marital status were all determined to be independent variables in the current data on the borrowers. The borrower's economic sector and the kind of loan that was disbursed are also independent variables. The variables stated above were chosen based on the completeness of the data acquired on them because the financial data obtained was rife with missing data. This answers the study's opening query.

Relevant theoretical and empirical literature were reviewed to inform the study, and gaps were identified. The following gaps were discovered during the literature review:

1. Little attention has been paid to the bank-based financial sectors in Sub-Saharan Africa and other developing nations, with the majority of credit scoring research concentrated on industrialised economies and large emerging markets to a lesser extent. Sub-Saharan Africa is the only region in the world without fully functional credit bureaus, with the exception of South Africa.
2. The best technique for developing a credit score model for all circumstances does not yet exist, however, logistic regression has been found to yield about sixty to eighty percent accuracy.
3. Most research use the 0.5 cut-off point as the scoring model's decision point without considering the imbalance nature of the data and investigating its effect on the model or the intended outcome.

The literature, however, confirmed that the wide gap in the measure for sensitivity and specificity is mainly due to the imbalance in data especially, financial data on loan defaulters and loan non-defaulters.

The population under study are individual borrowers of commercial banks in Ghana, however, a total of 8,687 borrowers were obtained from a commercial bank rather than from all twenty-three licenced commercial banks in Ghana. This sample of borrowers are considered to be representative of the population under study because all licensed commercial banks in Ghana use the same KYC policy prescribed by the central bank to gather information on customers. Using the Bernoulli random numbers generator in SPSS, the data was separated into 70% of borrowers that was used to develop the logistic model and the remaining 30% borrowers used to validate the model.

The logistic regression model was developed using IBM SPSS Modeller. The application revealed that the variables, tenor, interest rate, age, gender, and loan description are statistically significant at an $\alpha -$ level of 5%. It also revealed that at the even chance of occurrence (cut-off point) of 0.5, the sensitivity and specificity are 22.7% and 98.9% respectively, with a stated overall model performance of 93.3%. The model performed similarly on the 30% holdout sample of borrowers with sensitivity and specificity being 98.6% and 22.7% respectively with overall performance slightly diminished at 92.6%.

Diagnostic tests of the model revealed that the model fits the data. The ROC curve analysis conducted confirmed this fact by posting an accuracy measure of 91%.

Subsequently, a plot of sensitivity and specificity vs predicted probability from the model helped to derive the optimal cut-off point. The sensitivity and specificity curves' intersection showed that 0.087 was the optimal cut-off value for the model. The model's classification of true positives (sensitivity) improved from 22.7% to 85.5% at the optimal cut-off point, while its classification of true negatives (specificity) declined from 98.9% to a respectable 85.4%. The overall model performance also dropped slightly from 93.3% to 85.4%. Comparing other binary models' performance with the logistic regression derived at the optimal cut-off point showed similar results.

Finally, the study has established that the shifting of the cut-off point has no impact on the accuracy of the model. However, it is important to note that the cut-off point directly influences the number of prospective customers that qualify for credit, as well as the associated credit risk. In particular, increasing the cut-off point results in a larger pool of customers who are eligible for credit but also increases the potential loss due to credit risk.

Conclusion

Concerning the first objective, an examination of the data retrieved from the bank revealed that, in the name of customer turnaround time, the data collected on the borrowers is generally sufficient to satisfy the minimum KYC policy prescribed by the regulator, BOG. In addition, other information on the borrower, such as the domestic/residential address, mobile number, employer name, number of years with current employer, number of years at current residence, and number of dependants, were largely missing. Data entry omissions were identified in the data set, which was analysed and subsequently removed from the analysis. The logistic regression model developed to fit the data was found to be statistically significant, with the amount financed, interest rate, age, gender, and loan description as the significant independent variables. The binary logistic regression technique used to model the data adequately fits the data with an AUC accuracy of 91%, with a 95% confidence interval using the 0.5 cut-off point, confirming the assertions of (Abdou & Pointon, 2011). The obtained sensitivity and specificity of 22.7% and 98.9%, respectively are indicative of a typical unbalanced data set as discussed by (Hand & Henley, 1997a; Liang et al., 2020; Wang et al., 2021). The model was validated using the test set of data. The model correctly classified the test set with an accuracy of 92.6%.

On the second and second objective, the optimal cut-off point was determined to be 0.087, which indicated an accuracy measure for both sensitivity and specificity of 85% for the model obtained for the classification of loan defaulters and loan non-defaulters. This cut-off value was used to classify the borrowers into two groups: loan defaulters and loan non-defaulters. The overall accuracy of this classification was 85.4%. The performance of the logistic regression was found to be similar to other binary models.

Recommendations

Almost all of a company's actions rely on data, which serves as the foundation for decision-making at both the operational and strategic levels. Thus, inferior quality data may have a substantial detrimental influence on the efficiency of an organisation, whereas decent quality data are frequently critical to a company's long-term survival and growth (Batini et al., 2009;

Haug et al., 2011). Accurate and dependable data are necessary prerequisites for performing effective research. To ensure the data's quality, the data collecting process must be well-designed and the data must be gathered methodically.

From the findings, banks in Ghana have robust core banking application software that can compete with any of the developed economies in the world. These core banking applications are used to collate data on customers of the bank, mostly at the point of customer onboarding. However, data entry errors are quite common. Some entries are skipped, left blank, or completed with inaccurate information. For example, a mobile phone number is provided in the domicile address section of the entry form, a landline is entered in a form designated for mobile phone numbers, a wrong date of birth is captured, etc. This presents challenges for any rigorous analysis of the data for strategic business planning or evidence-based policy formulation that the bank can leverage against competition as discussed by (Even & Shankaranarayanan, 2009; Haug et al., 2011; Xiao et al., 2009). Realising the importance of quality data, some banks in Ghana have recently begun an in-house data cleaning and update of records exercise to sanitise their database.

Given the foregoing findings, this research makes the following recommendations: The responsibility of managing and maintaining data and databases resides with the IT department of the bank, therefore there should be top management engagement and interest in the quality and integrity of data. The IT department should be equipped with qualified staff that will ensure that the quality and integrity of the data collated are not compromised. There should be regular training of the data entry staff to ensure that they are consistent with the bank's established data import rules. There should be an implementation of a data cleaning routine to keep the data up to date and reduce data decay. Apart from the minimum KYC required by the regulator for customer onboarding, the bank may need to add a few fields such as the number of dependants and previous utilization of credit by the customer which are significant variables in credit scoring (Balina et al., 2021; Crook et al., 2007).

Credit scoring models such as those created by (Altman, 1968; Hand & Henley, 1997; Lee & Sung-Chang, 2000; Lovie & Lovie, 1986; Siddiqi, 2017; Thomas et al., 2002; W. E. Henley, 1996; Zhang, 2015) are among the most effective implementations of statistical modelling in finance and banking, as evidenced by the rising number of scoring analysts employed in the sector. Credit scoring has played a critical role in enabling the remarkable increases in consumer credit over the past years. The use of credit scoring algorithms, which serve as an accurate and automated risk assessment tool, has allowed consumer credit lenders to successfully grow their lending portfolios (Thomas et al., 2002).

Local banks are encouraged to use credit scoring tool, as proposed in this study, in addition to the existing credit risk management framework. It is expected that a credit scoring tool will help banks to more accurately assess the creditworthiness of borrowers, resulting in a more efficient allocation of resources and helping to reduce risks associated with lending.

The bank can take advantage of the wide range of statistical techniques used to develop scoring models. Most of these statistical models, some of which are nonlinear, are suitable for the development of an efficient and effective credit scoring system that can be used for prediction purposes effectively and efficiently. Credit analysts, researchers, lenders, and computer software developers use statistical techniques such as logistic regression, weight of evidence measure, probit analysis, regression analysis, linear programming, K-nearest neighbour (KNN), Cox's proportional hazard model, discriminant analysis, support vector machines, decision trees, neural networks, genetic algorithms, and genetic programming to build credit scoring models. Despite the above-mentioned range of techniques, research has shown that the best statistical technique for developing credit scoring models does not yet exist, however, the logistic regression technique consistently yields the most accurate measures (Abdou & Pointon, 2011). In addition, most statistical packages have an inbuilt algorithm to use logistic regression on the appropriate data and apart from the slight difficulty in interpretability, they are relatively easy to implement and use.

In this study, the credit score of a prospective borrower is obtained in the logistic model developed by simply summing up the product of various numbers designated for the answers received on the various variables and the coefficients of the variables to have a total score. At a cut-off point of 0.5, the prospective borrower’s total score is compared with the 0.5 cut-off point score. If this score is below the cut-off point of 0.5, credit is awarded, otherwise, the prospective borrower is denied. This study has confirmed that at the cut-off point of 0.5, the type II error is quite high and proposed the use of an optimal cut-off point of 0.087. It was shown that the optimal cut-off point of 0.087 greatly reduced the type II error and improved the reduction of credit risk. Although the 0.087 cut-off point may seem too low a cut-off point, the number of borrowers that may be rejected is still low and the measure of accuracy is still high at about 86%. In addition, the bank can use the cut-off point of 0.087 as the lower limit and the cut-off point of 0.5 as the upper limit to assess prospective borrowers. In this case, the borrower can be automatically granted credit if the score is below the lower cut-off, whereas credit is automatically rejected if it exceeds the upper cut-off. If the score is between the two cut-off points, the credit officer re-evaluates the borrower based on the actual requirements, or credit history information is obtained, scored, and the points are added to the total score obtained. As a result of this approach, if this new score is above the higher cut-off, credit is denied, if not, credit is granted. Furthermore, the bank can decide, based on its risk appetite, to use the optimal cut-off point as the upper limit to reduce credit risk to its barest minimum or extend the cut-off point to any point between 0.087 to 0.5 to expand the customer base. For any point above the 0.5 cut-off point, the bank might as well grant credit to any borrower that applies for it.

It is worth noting that the study's computed optimal cut-off point of 0.087 is specific to the local-controlled bank from which the data for computation was retrieved. Hence, the optimal cut-off point will differ from bank to bank depending on the number of variables, the sample size, and the risk appetite of the bank.

Areas for Further Research

There were two limitations worthy of note identified in the study. Firstly, the lack of access to data from all banks. Access to the other banks' borrowing customers would have enhanced the quality of this study. Future researchers that get access to this data can compare local-controlled banks' models to that of the foreign-controlled banks at various cut-off points. Secondly, the number of variables considered in this study were limited by data entry errors. As the data cleaning exercise is ongoing in some banks in Ghana, future researchers may get access to quality data for research. Access to quality data from all the banks in Ghana will enable the researcher to examine other variables that were not considered in this study.

REFERENCES

Abdou, H. A., & Pointon, J. (2011). Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2–3), 59–88, https://doi.org/10.1002/isaf.325.

Adedapo Soymeal, K., Jumoke Ogunleye, O., & Oluwatoyin Ashogbon, F. (2014). Risk Management Practices and Financial Performance: Evidence from the Nigerian Deposit Money Banks (DMBS). *International Journal of Managerial Studies and Research*, *2*(5), 31–39.

Afande, F. O. (2014). Credit Risk Management Practices of Commercial Banks in Kenya. *European Journal of Business and Management*, *6*(34), 21–62. www.iiste.org.

Akosa, J. S. (2017). Predictive accuracy: A misleading performance measure for highly imbalanced data. *SAS Global Forum*, *942*, 1–12.

Ali, A., Al-Turjman, F., & Al-Qutayri, M. (2020). Deep learning models for handling imbalanced data in computer vision applications: A review. *Journal of Systems and Software*, 170, 110702.

Allen, L., & Saunders, A. (2012). Risk Management in Banking. In *The Oxford Handbook of Banking* (p. 821). John Wiley, https://doi.org/10.1093/oxfordhb/9780199640935.013.0004.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589–609, https://doi.org/10.1111/j.1540-6261.1968.tb00843.x.

Altman, E. I. (1993). Corporate Financial Distress and Bankruptcy: A complete guide to predicting & avoiding distress and profiting from bankruptcy. In *Wiley-finance* (2nd ed.), Wiley and Sons.

Altman, E. I., & Saunders, A. (1996). Credit Risk Measurement Developments. In *New York University Stern School of Business Finance Department Working Paper Series*.

Amu, N. J. (2005). *The Role of Women in Ghana's Economy*. Friedrich Ebert Foundation.

Anderson, P., & Abousleiman, I. (2011). Risk Management Framework and Financial Solutions. *Presentation*.

Andersson, F., Mausser, H., Rosen, D., & Uryasev, S. (2000). Credit risk optimization with Conditional Value-at-Risk criterion. *Mathematical Programming, Series B*, *89*(2), 273–291, https://doi.org/10.1007/PL00011399.

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2016). Learning to learn with probing functions for online one-class classification. *In Proceedings of the 33rd International Conference on Machine Learning*, pp. 1633-1642.

Apanga, M. A.-N., Appiah, K. O., & Arthur, J. (2016). Credit Risk Management of Ghanaian Listed Banks. *International Journal of Law and Management*, *58*(2), 162–178. https://doi.org/10.1108/IJLMA-04-2014-0033.

Bachmair, F. F. (2016). Contingent Liabilities Risk Management: A Credit Risk Analysis Framework for Sovereign Guarantees and On-Lending: Country Experiences from Colombia, Indonesia, Sweden, and Turkey.

In *World Bank Policy Research Working Paper Series*. World Bank, Washington, DC. https://doi.org/10.1596/1813-9450-7538.

Badar, M., & Yasmin Javid, A. (2013). Impact of macroeconomic forces on nonperforming loans: An empirical study of commercial banks in Pakistan. *WSEAS Transactions on Business and Economics*, *10*(1), 40–48.

Balina, R., Idasz-Balina, M., Ortolano, A., & Febo, E. di. (2021). Drivers of Individual Credit Risk of Retail Customer: A Case Study on the Example of the Polish Cooperative Banking Sector. *Risks 2021, Vol. 9, Page 219*, *9*(12), 219, https://doi.org/10.3390/RISKS9120219.

Bank of Ghana. (2017). *Notice to the Banks and The General Public Notice No. BG/GOV/SEC/2017/19 New Minimum Paid Up Capital*. https://www.bog.gov.gh/wp-content/uploads/2019/09/Notice-on-New-Minimum-Paid-Up-Capital.pdf.

Bank of Ghana. (2018). *Press Release – Government Establishes New Indigenous Bank*. https://www.bog.gov.gh/wp-content/uploads/ 2019/07/PRESS-RELEASE-Grand-Final-August-2018.pdf.

Bank of Ghana. (2019). Notice of revocation of licences of insolvent savings and loans companies and finance houses, and appointment of a receiver. *Press Release*, *2008*(Act 774), 1–27.

Bank of Ghana. (n.d.). *Know Your Customer Policy – Bank of Ghana*. Retrieved February 17, 2021, from https://www.bog.gov.gh/supervision-regulation/know-your-customer-policy/.

Barua, S., Islam, M. M., & Yao, X. (2014). MWMOTE—Majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2), 405-425.

Basel Committee on Banking Supervision. (2005). *Basel Committee on Banking Supervision, An Explanatory Note on the Basel II IRB Risk Weight Functions*.

Basel Committee on Banking Supervision. (2016). *Basel Committee on Banking Supervision*.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, *41*(3). https://doi.org /10.1145/ 1541880. 1541883.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29.

Bellini, T. (2011). *Stress Testing and Risk Integration in Banks: A statistical framework and practical software guide (in MATLAB and R)* (third edition). Wiley. https://www.wiley.com/en-us/9780470689851.

Bensusan, H., Malerba, D., & Semeraro, G. (2000). A hierarchical stacking method for learning from heterogeneous sources. *Machine Learning*, 40(2), 151-175.

Bernanke, B. S. (1993). Credit in the Macroeconomy. *Quarterly Review*, *18*(Spr), 50–70.

Bessis, J. (2002). *Risk Management in Banking* (Second). John Wiley & Sons, Ltd. https://www.mendeley.com/catalogue/b1763198-ff6f-3ce6-8649-428f443fdd1b/?utm_source=desktop&utm_medium=1.19.8&utm_campaign=open_catalog&userDocumentId=%7B5b104e9a-b4e3-4432-ab96-7e1a273218f4%7D.

Bettis, R. A., & Howard, T. (1990). Risk, Strategy, and Management. *JAI Press*, *24*(5), 215. https://doi.org/https://doi.org/10.1016/0024-6301(91)90264-O

Bhardwaj, T. (2013). Credit risk management practices: A comparative study of state of India and Punjab National Bank. *Banking Sector Report*, *2*(3), 23–45.

Bloom, M., & Milkovich, G. T. (1998). Relationships among Risk, Incentive Pay, and Organizational Performance. *Academy of Management Journal*, *41*(3), 283–297. https://doi.org/10.2307/256908.

Boahene, S. H., Dasah, J., & Agyei, S. K. (2012). Credit Risk and Profitability of Selected Banks in Ghana. *Research Journal of Finance and Accounting*, *3*(7), 6–14.

BOG, & FIC. (2018). *Bank of Ghana and Financial Intelligence Centre Anti-Money Laundering/ Combating the Financing of Terrorism & The Proliferation of Weapons of Mass Destruction (AML/CFT& P) Guideline*.

Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the Independent Variables. *Technometrics*, *4*(4), 531. https://doi.org/10.2307/1266288

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Bulmer, M. G. (1979). *Principles of statistics* (Dover). Dover Publications Inc.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. *In Advances in Knowledge Discovery and Data Mining* (pp. 475-482). Springer.

Campbell, A. (2007). Bank Insolvency and the Problem of Non-Performing Loans. *International Insolvency Law: Themes and Perspectives*, 211–235. https://doi.org/10.1057/PALGRAVE.JBR.2350057.

Carey, M., & Stulz, R. M. (2005). The Risks of Financial Institutions. *NBER Working Papers*.

Cebenoyan, A. S., & Strahan, P. E. (2004). Risk Management, Capital Structure and Lending at Banks. *Journal of Banking and Finance, 28*(1), 19–43. https://doi.org/10.1016/S0378-4266(02)00391-6.

Cebotari, A. (2008). Contingent Liabilities: Issues and Practice. *IMF Working Papers*, *08*(245), 1–60. https://doi.org/10.5089/9781451871036.001.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1-58.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: improving prediction of the minority class in boosting. *In European conference on principles of data mining and knowledge discovery* (pp. 107-119). Springer.

Clavel, J., Merceron, G., & Escarguel, G. (2014). Missing Data Estimation in Morphometrics: How Much is Too Much? *Systematic Biology*, *63*(2), 203–218. https://doi.org/10.1093/SYSBIO/SYT100.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent Developments in Consumer Credit Risk Assessment. *European Journal of Operational Research*, *183*(3), 1447–1465. https://doi.org /10.1016/ j.ejor.2006.09.100.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *In Proceedings of the 23rd International Conference on Machine Learning*, pp. 233-240.

Dietrich, A., & Wanzenried, G. (2011). Determinants of Bank Profitability Before and During the Crisis: Evidence From Switzerland. *Journal of International Financial Markets, Institutions and Money*, *21*(3), 307–327. https://doi.org/10.1016/J.INTFIN.2010.11.002.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman \& Hall/CRC.

Donaldson, L. (2000). Organizational Portfolio Theory: Performance-Driven Organizational Change. *Contemporary Economic Policy*, *18*(4), 386. https://doi.org/10.1111/j.1465-7287.2000.tb00035.x.

Douzas, G., & Bacao, F. (2019). G-SMOTE: a general framework for oversampling with geometric SMOTE. *Information Sciences*, 501, 378-397.

Einarsson, A. I. (2008). Credit Risk Modeling [Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark]. In *Technical University of Denmark*. https://doi.org/10.1142/9789813142114_0002.

Elijah, B. (2019). *Bank Failure in Ghana: What Accounted for the Collapse of Unibank?*

Elkan, C. (2001). The foundations of cost-sensitive learning. *In International joint conference on artificial intelligence* (pp. 973-978).

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*.

Even, A., & Shankaranarayanan, G. (2009). Utility Cost Perspectives in Data Quality Management. *Journal of Computer Information Systems*, *50*(2), 127–135. https://www.proquest.com/openview/_e9e95db_0faf0bd025478c9aa3d1f412c/1?pq-origsite=gscholar&cbl=48936.

Fabozzi, F. J., Mann, S. V., & Choudhry, M. (2003). Credit Risk Measuring and Controlling, Interest Rate and Credit Risk. In *John Wiley & Sons, Inc.* (Second). John Wiley & Sons, INC., Publication. https:// doi.org/10.1111/jmcb.12182.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.

Fernández, A., García, S., & Herrera, F. (2018). Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*, 22(11), 3859-3876.

García, A. M. V, & Sánchez, J. (2013). On the Suitability of Resampling Techniques for the Class Imbalance Problem in Credit Scoring. *The*

Journal of the Operational Research Society, *Vol 64*(No. 7), 1060–1070.

García, S., & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, 17(3), 275-306.

Gizycki, M., Ellis, L., Fisher, C., Gruen, D., Heath, A., Kent, C., Lowe, P., Simon, J., Voss, G., Mcmahon, A., & Razzaghipour, A. (2001). *The Effect of Macroeconomic Conditions on Banks' Risk and Profitability.* System Stability Department Reserve Bank of Australia.

Gup, B. E., & Kolari, J. W. (2005). *Commercial Banking: The Management of Risk* (Third). J. Wiley.

Hand, D. J., & Henley, W. E. (1997a). Statistical classification methods in consumer credit scoring: a review. *J Roy Stat Soc A*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985x.1997.00078.x.

Hand, D. J., & Henley, W. E. (1997b). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x.

Hannan, T. H., & Hanweck, G. A. (1988). Bank Insolvency Risk and the Market for Large Certificates of Deposit. *Journal of Money, Credit and Banking*, *20*(2), 203. https://doi.org/10.2307/1992111.

Harris, T. (2013). Default Definition Selection for Credit Scoring. *Artificial Intelligence Research*, *2*(4), 49–62. https://doi.org/ 10.5430/ air.v 2n4 p49.

Haug, A., Zachariassen, F., van Liempd, D., Haug, A., Zachariassen, F., van Liempd, D., Haug, A., & Zachariassen, F. (2011). The Costs of Poor Data Quality. *Journal of Industrial Engineering and Management (JIEM)*, *4*(2), 168–193. https://doi.org/ 10.3926 /JIEM.2011. V4N2.P168-193.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In International Joint Conference on Neural Networks* (pp. 1322-1328). IEEE.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: adaptive synthetic sampling approach for imbalanced learning. *In International joint conference on neural networks* (pp. 1322-1328). IEEE.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832-844.

Hosmer, D. W. Jr., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. In D. J. Balding, N. A. C. Cressie, G. M. Fitzmaurice, H. Goldstein, I. M. Johnstone, G. MolenberGH¢, D. W. Scott, A. F. M. Smith, R. S. Tsay, S. Weisberg, V. Barnett, J. S. Hunter, J. B. Kadane, & J. L. Teugels (Eds.), *John Wiley & Sons, Inc.* (Third, Vol. 40, Issue 4). John Wiley & Sons, INC., Publication. https://doi.org/10.2307/2348743.

Jackson, P., & Perraudin, W. (2000). Regulatory Implications of Credit Risk Modelling. *Journal of Banking and Finance*, *24*(1–2), 1–14. https://doi.org/10.1016/S0378-4266(99)00050-3.

Jafari, M., Chadegani, A. A., & Biglari, V. (2011). Effective Risk Management and Company's Performance: Investment in Innovations and Intellectual Capital using Behavioral and Practical Approach. *Journal of Economics and International Finance*, *3*(15), 780–786. https://doi.org/10.5897/JEIF11.123.

Jiang, B., Li, X., & Zhang, C. (2018). A deep learning approach to classifying imbalanced fault diagnosis data. Measurement, 128, 98-107.

Jonathan, P. (2012). Credit Risk Management in Banking Industry: Case Study Atwiman Kwanwoma Rural Bank By. *Risk Manag RM377*, 109.

Kanchu, T., & Kumar, M. M. (2013). Risk Management in Banking Sector-An Empirical Study. *International Journal of Marketing, Financial Services & Management Research*, *2*(2).

Khan, M. S., & Madden, M. G. (2014). A survey of recent trends in one class classification. *In Proceedings of the 10th International Conference on Signal-Image Technology & Internet-Based Systems*, pp. 8-15.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer Credit-Risk Models Via Machine-Learning Algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787. https://doi.org/ 10.1016/ J.JBANKFIN. 2010.06.001.

Kiselakova, D., Horváthová, J., Sofrankova, B., & Šoltés, M. (2015). Analysis Of Risks and their Impact on Enterprise Performance by Creating Enterprise Risk Model. *Polish Journal of Management Studies*, *11*, 50–61.

Kiselakova, Dana., & Kiselak, A. (2013). Analysis of Banking Business and its Impact on Financial Stability of Economies in Euro Area. *Polish Journal of Management Studies, Volume Vol. 8*.

Kleimeier, S., & Dinh, T. (2007). A Credit Scoring Model for Vietnam's Retail Banking Market. *International Review of Financial Analysis, 16*(5), 471–495.

Koch, T. W., & Scott, M. S. (2014). Bank Management. In *Cengage Learning* (Vol. 23, Issue 2). https://doi.org/10.1108/eb002056.

Koford, K., & Tschoegl, A. E. (1999). Problems of Bank Lending in Bulgaria: Information Asymmetry and Institutional Learning. *MOST: Economic Policy in Transitional Economies 1999 9:2*, *9*(2), 123–152. https://doi.org/10.1023/A:1009508611388.

Konovalova, N., Kristovska, I., & Kudinska, M. (2016). Credit Risk Management in Commercial Banks. *Polish Journal of Management Studies*, *13*(2), 90–100.

Kun, Q., & Duo, M. (2014a). Credit Risk Management of a Commercial Bank. *Journal of Chemical and Pharmaceutical Research*, *6*(5), 1784–1788. https://doi.org/10.18411/spc-08-11-2018-01.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. *Technical report, Tampere University of Technology*.

le Bellac, M., & Viricel, A. (2016). *Deep Dive into Financial Models: Modeling Risk and Uncertainty*. Co-published With World Scientific.

Lee, T. H., & Sung-Chang, J. (2000). Forecasting Creditworthiness: Logistic vs. Artificial Neural Network. *The Journal of Business Forecasting Methods & Systems*, *18*(4), 28–30. https://search.proquest.com/openview/fa76a651a071e77ed6472b3cb81360d3/1?pq-origsite=gscholar&cbl=28144.

Lewis, C. M., & Mody, A. (1997). The Management of Contingent Liabilities: A Risk Management Framework for National Governments. In T. Irwin, M. Klein, G. E. Perry, & M. Thobani (Eds.), *World Bank latin America and Caribbean Studies*. World Bank, Washington, DC.

Li, W., Zhang, X., & Liu, X. (2019). A review of instance selection methods for imbalanced data sets. *Artificial Intelligence Review*, 52(1), 43-58.

Liang, X. W., Jiang, A. P., Li, T., Xue, Y. Y., & Wang, G. T. (2020). LR-SMOTE — An Improved Unbalanced Data Set Oversampling Based on K-Means and SVM. *Knowledge-Based Systems*, *196*, https://doi.org/10.1016/j.knosys.2020.105845.

Liu, B., Wang, L., & Yu, S. (2009). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 23(1), 49-57.

Liu, Y., Zhang, J., Sun, G., & Cheng, X. (2021). An overview of oversampling and undersampling in data pre-processing for machine learning. *Big Data Mining and Analytics*, 4(4), 254-279.

Lopez, J. A., & Saidenberg, M. R. (1999). *Evaluating Credit Risk Models*.

Lovie, A. D., & Lovie, P. (1986). The Flat Maximum Effect and Linear Scoring Models for Prediction. *Journal of Forecasting*, *5*(3), 159–168. https://doi.org/10.1002/for.3980050303.

Matthews, K., & Thompson, J. (2020). The Economics of Banking. In *The Economics of Banking*. https://doi.org/10.4337/9781789902761.

Memic, D. (2015). Assessing Credit Default using Logistic Regression and Multiple Discriminant Analysis: Empirical Evidence from Bosnia and Herzegovina. *Interdisciplinary Description of Complex Systems*, *13*(1), 128–153. https://doi.org/10.7906/indecs.13.1.13.

Merton, R. C., & Perold, A. (1993). Theory of Risk Capital in Financial Firms. *Journal of Applied Corporate Finance*, *6*(3), 16–32, https://doi.org/10.1111/J.1745-6622.1993.TB00231.X.

Montiel, P., & Servén, L. (2006). Macroeconomic Stability in Developing Countries: How much is enough? In *World Bank Research Observer* (Vol. 21, Issue 2, pp. 151–178). https://doi.org/10.1093/wbro/lkl005.

Moore, P. G., & Rao, C. R. (1967). Linear Statistical Inference and its Applications. In *The Statistician* (Vol. 17, Issue 1), https://doi.org/10.2307/2987205.

Moreno, R. (2006). *The changing nature of risks facing banks*. *28*(4), 67–98.

Mozib Lalon, R. (2015). Credit Risk Management (CRM) Practices in Commercial Banks of Bangladesh: “A Study on Basic Bank Ltd.” *International Journal of Economics, Finance and Management Sciences*, *3*(2), 78, https://doi.org/10.11648/j.ijefm.20150302.12.

Naimy, V. Y. (2005). Overall Lebanese Banks Performance: A Risk-Return Framework. *International Business & Economics Research Journal (IBER)*, *4*(1). https://doi.org/10.19030/IBER.V4I1.3564.

Nocco, B. W., & Stulz, R. M. (2006). Enterprise Risk Management: Theory and Practice. *Journal of Applied Corporate Finance*, *18*(4), 8–20. https://doi.org/10.1111/J.1745-6622.2006.00106.X.

Odonkor, T. A., Osei, K. A., Abor, J., & Adjasi, C. K. D. (2011). Bank Risk and Performance in Ghana. *International Journal of Financial Services Management*, *5*(2), 107–120. https://doi.org/ 10.1504/IJFSM. 2011. 041919.

Ofosu-Hene, E. D., & Amoh, P. (2016). Risk Management and Performance of Listed Banks in Ghana. *European Journal of Business Science and Technology*, *2*(2), 107–121. https://doi.org/10.11118/ejobsat.v2i2.46.

Oldfield, G. S., & Santomero, A. M. (1997). The Place of Risk Management in Financial Institutions. *Sloan Management Review*, *39 No.1*, 33–46.

Onang'o, O. N. (2017). *Effect of Credit Risk Management on Financial Performance of Commercial Banks Listed at the Nairobi Securities Exchange, Kenya*. Kenyatta University.

Pang, W., Liu, Y., & Yuan, C. (2020). Efficient oversampling techniques for imbalanced data sets. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4), 1419-1431.

Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12), 3448-3470.

Pituch, K. A., & Stevens, J. P. (2015). Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM's SPSS, Sixth Edition. In *Applied Multivariate Statistics for the Social Sciences: Analyses* *with SAS and IBM's SPSS, Sixth Edition*. https://doi.org/ 10.4324/ 9781315814919.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3), 21-45.

Poudel, R. P. S. (2012). The Impact of Credit Risk Management on Financial Performance of Commercial Banks in Nepal. *International Journal of Arts and Commerce*, *1*(5).

Pregibon, D. (1980). Goodness of Link Tests for Generalized Linear Models. *Applied Statistics*, *29*(1), 15. https://doi.org/10.2307/2346405.

PricewaterhouseCoopers. (2019). *Banking Reforms So Far: Topmost Issues on the Minds of Bank CEOs* (Issue August).

Rawlings, J. O., Pantula, S. G., & Dickey, D. a. (1998). *Applied Regression Analysis: A Research Tool* (Second Edi). Springer.

Sakyi, P. A., Ofoeda, I., Kyereboah-Coleman, A., & Abor, J. Y. (2014). Risk and Performance of Non-Bank Financial Institutions. *Int. J. Financial Services Management*, *7*(1), 19–35.

Saunders, A., & Cornett, M. M. (2008a). Financial Institutions Management: A Risk Management Approach. In *Financial Institutions Management* (Sixth). McGraw-Hill.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2), 197-227.

Schreiner, M. (2003). Scoring: The Next Breakthrough in Microcredit? *CGAP, World Bank, 7*.

Schroeck, G. (2002). *Risk Management and Value Creation in Financial Institutions*. John Wiley & Sons, INC., Publication.

Siddiqi, N. (2017). Intelligent Credit Scoring. In *Intelligent Credit Scoring*. https://doi.org/10.1002/9781119282396.

Smithson, C., & Simkins, B. J. (2005). Does Risk Management Add Value? A Survey of the Evidence. *Journal of Applied Corporate Finance*, *17*(3), 8–17. https://doi.org/10.1111/J.1745-6622.2005.00042.X

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.

Sun, Y., Wong, A. K., & Kamel, M. S. (2016). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 30(12), 1-35.

Tan, Y., Wang, D., & Jiang, W. (2020). Deep learning for imbalanced fault classification based on a modified oversampling approach. *IEEE Access*, 8, 35846-35857.

Tandelilin, E., Kaaro, H., Mahadwartha, P. A., & Supriyatna. (2007). Corporate Governance, Risk Management, and Bank Performance: Does Type of Ownership Matter? *EADN Working Paper*, *34*, 1–141.

Tax, D. M. J., & Duin, R. P. W. (1999). Support vector data description. *Machine Learning*, 54(1), 45-66.

The Basel Committee. (2000). Principles for the Management of Credit Risk. *IFAS Extension*, 1–33.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit Scoring and its Application*. Society for Industrial and Applied Mathematics.

Turner, P. (2006). The Banking System in Emerging Economies: How Much Progress Has Been Made? *BIS Papers*, *28*, 1–11.

U.S. Small Business Administration. (2005). *U.S. SBA-Office of Advocacy--The Small Business Economy; A Report to the President, 2005*. www.sba.gov/advo.

van Greuning, H., & Brajovic Bratanovic, S. (2009). Analyzing Banking Risk (3rd Edition). In *Analyzing Banking Risk (3rd Edition)*. The World Bank. https://doi.org/10.1596/978-0-8213-7728-4.

Vogelgesang, U. (2003). Microfinance in Times of Crisis: The Effects of Competition, Rising Indebtness, and Economic Crisis on Repayment Behaviour. *World Development*, *31*(12), 2085–2114.

W. E. Henley. (1996). Nearest neighbour analysis in credit scoring.pdf. In *The Statistician: Vol. Volume 45* (Issue 1, pp. 77–95).

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3074243.

Wang, X., Chen, C., Li, L., & Liu, Z. (2019). A survey of transfer learning for imbalanced data processing. *Neurocomputing*, 365, 71-91.

Weisberg, S. (2005). *Applied Linear Regression* (Third). Wiley and Sons.

Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. *Morgan Kaufmann Publishers*.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241-259.

Xiao, J. H., Xie, K., & Wan, X. W. (2009). Factors Influencing Enterprise to Improve Data Quality in Information Systems Application - An Empirical Research on 185 Enterprises Through Field Study. *2009*

International Conference on Management Science and Engineering - 16th Annual Conference Proceedings, ICMSE 2009, 23–33. https://doi.org/10.1109/ICMSE.2009.5317534.

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, *2*(3), 249–262. https://doi.org/10.1007/S41664-018-0068-2/FIGURES/9.

Xu, Y., Xie, X., & Wang, Z. (2018). An improved KNN algorithm for classifying imbalanced data. *Applied Soft Computing*, 68, 797-807.

Yadolah, D. (2008). The Concise Encyclopedia of Statistics. In *Springer*.

Zhang, Q. (2015). *Modeling the Probability of Mortgage Default Via Logistic Regression and Survival Analysis*. http://digitalcommons.uri.edu/theseshttp://digitalcommons.uri.edu/theses/541.

Zhang, X. Y., Yin, Y. H., & Cui, L. Z. (2017). A novel resampling method for imbalanced data based on Euclidean distance distribution. *Journal of Intelligent Information Systems*, 48(2), 339-354.

Zhang, X., Li, W., & Liu, X. (2018). A survey of deep learning-based approaches for handling imbalanced data sets. *Neural Computing and Applications*, 30(6), 1623-1637.

APPENDIX A

DECRIPTIVE STATISTICS FOR THE CATEGORICAL VARIABLES

Descriptive Statistics for Loan Default

	Frequency	Percent	Valid Percent	Cumulative Percent
NO	8,035	92.5	92.5	92.5
YES	652	7.5	7.5	100.0
Total	8,687	100.0	100.0	

Descriptive Statistics for Gender

	Frequency	Percent	Valid Percent	Cumulative Percent
FEMALE	3,281	37.8	37.8	37.8
MALE	5,406	62.2	62.2	100.0
Total	8,687	100.0	100.0	

Descriptive Statistics for Marital Status

	Frequency	Percent	Valid Percent	Cumulative Percent
DIVORCED	326	3.8	3.8	3.8
ENGAGED	396	4.6	4.6	8.3
MARRIED	3,929	45.2	45.2	53.5
SINGLE	3,912	45.0	45.0	98.6
WIDOWED	124	1.4	1.4	100.0
Total	8,687	100.0	100.0	

Descriptive Statistics for Employment Type

	Frequency	Percent	Valid Percent	Cumulative Percent
FULL TIME	8,375	96.4	96.4	96.4
PART TIME	16	.2	.2	96.6
RETIRED	241	2.8	2.8	99.4
OTHER	55	.6	.6	100.0
Total	8,687	100.0	100.0	

Descriptive Statistics for Loan Description/type

	Frequency	Percent	Valid Percent	Cumulative Percent
CAGD SALARY LOAN SCHEME	4,300	49.5	49.5	49.5
COMMERCIAL LOANS	256	2.9	2.9	52.4
INSTITUTIONAL WORKERS LOAN SCHEME	2,862	32.9	32.9	85.4
MICRO CREDIT LOAN SCHEME	274	3.2	3.2	88.5
SALARY WORKERS LOAN SCHEME	995	11.5	11.5	100.0
Total	8,687	100.0	100.0	

Descriptive Statistics for Economic Sector

	Frequency	Percent	Valid Percent	Cumulative Percent
PRIMARY	26	0.3	0.3	0.3
TERTIARY	8,582	98.8	98.8	99.1
SECONDARY	79	.9	.9	100.0
Total	8,687	100.0	100.0	

APPENDIX B

DESCRIPTIVE STATISTICS FOR THE NUMERICAL VARIABLES

Descriptive Statistics for the Numerical Variables

	Amount Financed (GH¢)	Tenor (Months)	Interest Rate (%)	Age (Years)
N	8,687	8,687	8,687	8,687
Mean	13,235.73	1,136.89	27.48	39.88
Median	8,217.60	1,094	28	39
Mode	5,127	1,094	28	37
Skewness	6.10	-0.24	1.45	0.27
Kurtosis	75.504	-0.099	5.553	-0.705
Range	469,624.85	2,524	26	63
Minimum	550.59	30	20	20
Maximum	470,175.44	2,554	46	83

APPENDIX C

HISTOGRAM OF NUMERICAL VARIABLES

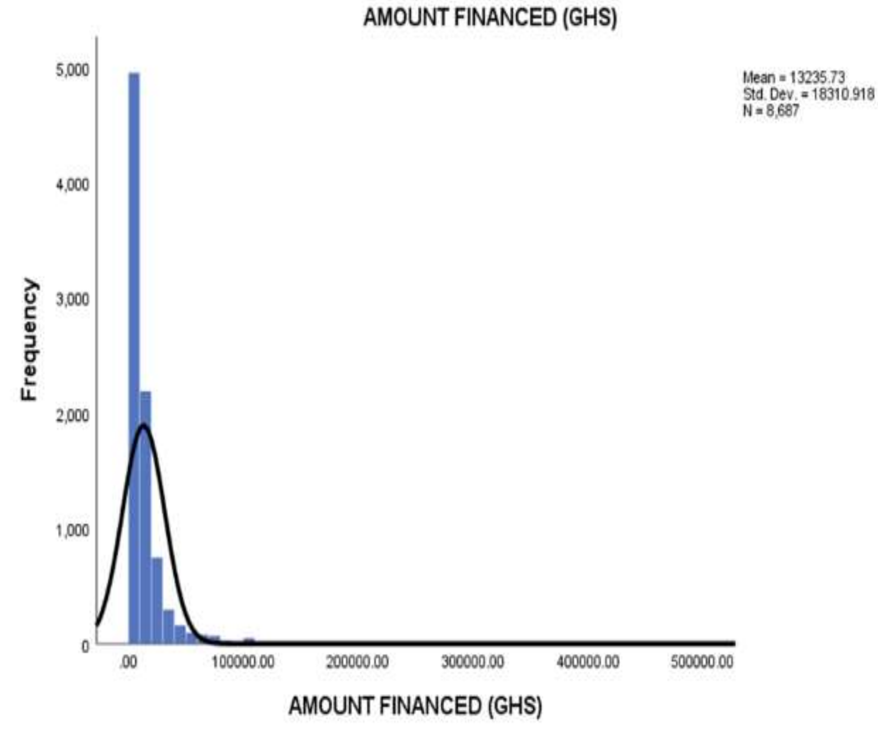

Figure 10. Distribution of Amount Financed

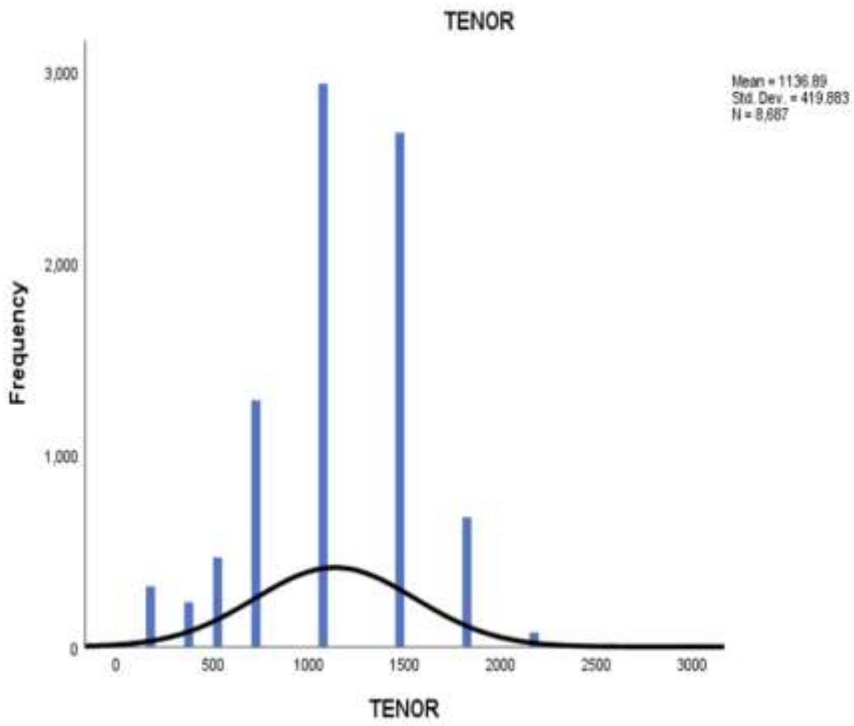

Figure 11. Distribution of Tenor

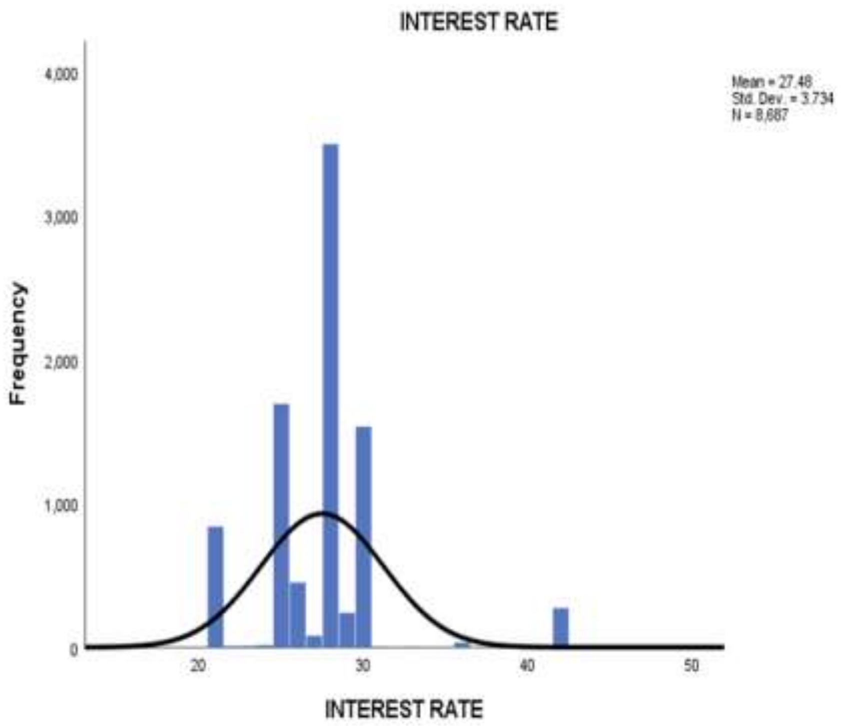

Figure 12. Distribution for Interest Rate

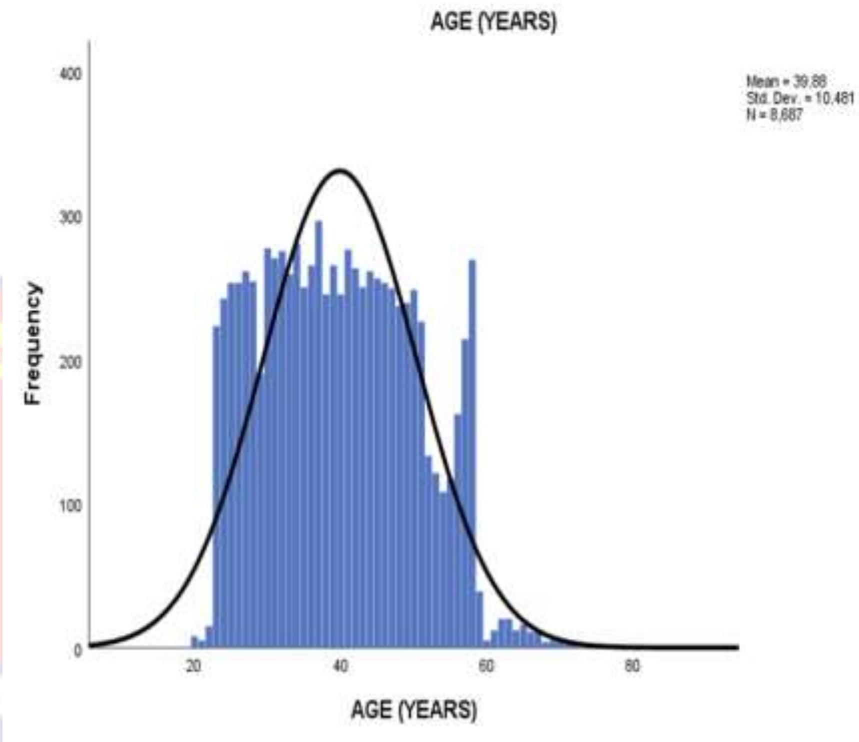

Figure 13. Distribution for Age

APPENDIX D
TEST FOR MULTICOLLINEARITY

Spearman's Correlation

		AMOUNT FINANCED	TENOR	INTEREST RATE	AGE	GENDER	MARITAL STATUS	TYPE OF EMPLOYMENT	LOAN DESCRIPTION	ECONOMIC SECTOR
AMOUNT FINANCED	Correlation Coefficient	1.000	.571**	-.293**	.042**	.161**	-.054**	-.096**	.035**	-.078**
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000	.001	.000
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687
TENOR	Correlation Coefficient	.571**	1.000	-.356**	.002	.181**	-.053**	-.245**	-.051**	-.097**
	Sig. (2-tailed)	.000	.	.000	.824	.000	.000	.000	.000	.000
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687
INTEREST RATE	Correlation Coefficient	-.293**	-.356**	1.000	-.045**	-.303**	.064**	.194**	-.120**	.067**
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000	.000	.000	.000
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687
AGE	Correlation Coefficient	.042**	.002	-.045**	1.000	.051**	.024*	.033**	.074**	-.013
	Sig. (2-tailed)	.000	.824	.000	.	.000	.028	.002	.000	.238
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687
GENDER	Correlation Coefficient	.161**	.181**	-.303**	.051**	1.000	-.071**	-.038**	.436**	.020
	Sig. (2-tailed)	.000	.000	.000	.000	.	.000	.000	.000	.067
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687
MARITAL STATUS	Correlation Coefficient	-.054**	-.053**	.064**	.024*	-.071**	1.000	-.004	-.092**	.014
	Sig. (2-tailed)	.000	.000	.000	.028	.000	.	.708	.000	.197
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687
TYPE OF EMPLOYMENT	Correlation Coefficient	-.096**	-.245**	.194**	.033**	-.038**	-.004	1.000	.241**	.074**
	Sig. (2-tailed)	.000	.000	.000	.002	.000	.708	.	.000	.000
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687

		AMOUNT FINANCED	TENOR	INTEREST RATE	AGE	GENDER	MARITAL STATUS	TYPE OF EMPLOYMENT	LOAN DESCRIPTION	ECONOMIC SECTOR
LOAN DESCRIPTION	Correlation Coefficient	.035**	-.051**	-.120**	.074**	.436**	-.092**	.241**	1.000	.097**
	Sig. (2-tailed)	.001	.000	.000	.000	.000	.000	.000	.	.000
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687
ECONOMIC SECTOR	Correlation Coefficient	-.078**	-.097**	.067**	-.013	.020	.014	.074**	.097**	1.000
	Sig. (2-tailed)	.000	.000	.000	.238	.067	.197	.000	.000	.
	N	8687	8687	8687	8687	8687	8687	8687	8687	8687

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Results of the Box-Tidwell test

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Amount Financed	.000	.000	.001	1	.971	1.000	1.000	1.000
Tenor	-.012	.004	11.505	1	.001	.988	.981	.995
Interest Rate	-.180	.922	.038	1	.845	.835	.137	5.093
Age	.450	.143	9.940	1	.002	1.569	1.186	2.075
Gender(1)	.556	.129	18.718	1	.000	1.744	1.355	2.243
Marital Status			10.163	4	.038			
Marital Status(1)	.182	.259	.494	1	.482	1.200	.722	1.993
Marital Status(2)	.100	.212	.220	1	.639	1.105	.729	1.675
Marital Status(3)	-.201	.210	.911	1	.340	.818	.542	1.235
Marital Status(4)	.344	.363	.898	1	.343	1.411	.692	2.875
Type of Employment			6.773	3	.079			
Type of Employment(1)	.086	.674	.016	1	.898	1.090	.291	4.087
Type of Employment(2)	-.411	.177	5.379	1	.020	.663	.468	.938
Type of Employment(3)	.301	.307	.962	1	.327	1.351	.740	2.466
Loan Description			170.331	4	.000			
Loan Description(1)	18.928	603.884	.001	1	.975	166144174.557	.000	.
Loan Description(2)	18.511	603.884	.001	1	.976	109456421.132	.000	.
Loan Description(3)	15.582	603.884	.001	1	.979	5851895.447	.000	.
Loan Description(4)	20.152	603.884	.001	1	.973	565066231.984	.000	.
Economic Sector			.718	2	.698			
Economic Sector(1)	-.013	.493	.001	1	.979	.987	.376	2.593
Economic Sector(2)	-.235	.550	.182	1	.670	.791	.269	2.324
Amount Financed by LN_AMTFIN	.000	.000	.121	1	.728	1.000	1.000	1.000
LN_TENOR by Tenor	.001	.000	.397	1	.482	1.001	1.001	1.002
Interest Rate by LN_INTRATE	.091	.214	.179	1	.672	1.095	.720	1.664
Age by LN_AGE	-.097	.030	.386	1	.537	.907	.855	.963
Constant	-25.951	603.914	.002	1	.966	.000		

a. Variable(s) entered on step 1: Amount Financed, Tenor, Interest Rate, Age, Gender, Marital Status, Type of Employment, Loan Description, Economic Sector, Amount Financed * LN_AMTFIN , LN_TENOR * Tenor , Interest Rate * LN_INTRATE , Age * LN_AGE .

APPENDIX E

CONFUSION MATRICES FOR THE SELECTED BINARY ALGORITHMS

CHAID Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,585	18	**99.7%**	2,429	3	**99.9%**	8,014	21	**99.7%**
	Yes	393	62	**13.6%**	166	31	**15.7%**	559	93	**14.3%**
			% Overall	**93.2%**		**% Overall**	**93.6%**		**% Overall**	**93.3%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,813	819	**85.5%**	2,062	355	**85.3%**	6,875	1,174	**85.4%**
	Yes	628	5000	**88.8%**	247	2137	**89.6%**	875	7,137	**89.1%**
			% Overall	**87.1%**		**% Overall**	**87.5%**		**% Overall**	**87.2%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	412	58	**87.7%**	168	27	**86.2%**	580	85	**87.2%**
	Yes	44	418	**90.5%**	20	174	**89.7%**	64	592	**90.2%**
			% Overall	**89.1%**		**% Overall**	**87.9%**		**% Overall**	**88.7%**

Tree-AS Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,585	18	**99.7%**	2,429	3	**99.9%**	8,014	21	**99.7%**
	Yes	393	62	**13.6%**	166	31	**15.7%**	559	93	**14.3%**
			% Overall	**93.2%**		**% Overall**	**93.6%**		**% Overall**	**93.3%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,888	740	**86.9%**	2,098	306	**87.3%**	6,986	1,046	**87.0%**
	Yes	477	5,162	**91.5%**	204	2,200	**91.5%**	681	7,362	**91.5%**
			% Overall	**89.2%**		**% Overall**	**89.4%**		**% Overall**	**89.3%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	385	70	**84.6%**	178	36	**83.2%**	563	106	**84.2%**
	Yes	50	407	**89.1%**	24	163	**87.2%**	74	570	**88.5%**
			% Overall	**86.8%**		**% Overall**	**85.0%**		**% Overall**	**86.3%**

XGBoost Tree Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,603	0	**100.0%**	2,432	0	**100.0%**	8,035	0	**100.0%**
	Yes	455	0	**0.0%**	197	0	**0.0%**	652	0	**0.0%**
			% Overall	**92.5%**		**% Overall**	**92.5%**		**% Overall**	**92.5%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,593	0	**100.0%**	2,433	0	**100.0%**	8,026	0	**100.0%**
	Yes	5,667	0	**0.0%**	2,372	0	**0.0%**	8,039	0	**0.0%**
			% Overall	**49.7%**		**% Overall**	**50.6%**		**% Overall**	**50.0%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	447	0	**100.0%**	201	0	**100.0%**	648	0	**100.0%**
	Yes	454	0	**0.0%**	194	0	**0.0%**	648	0	**0.0%**
			% Overall	**49.6%**		**% Overall**	**50.9%**		**% Overall**	**50.0%**

Random Forest Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,598	5	**99.9%**	2,396	36	**98.5%**	7,994	41	**99.5%**
	Yes	24	431	**94.7%**	93	104	**52.8%**	117	535	**82.1%**
			% Overall	**99.5%**		**% Overall**	**95.1%**		**% Overall**	**98.2%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,581	65	**98.8%**	2,355	29	**98.8%**	7,936	94	**98.8%**
	Yes	0	5,631	**100.0%**	0	2,432	**100.0%**	0	8,063	**100.0%**
			% Overall	**99.4%**		**% Overall**	**99.4%**		**% Overall**	**99.4%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	374	69	**84.4%**	178	29	**86.0%**	552	98	**84.9%**
	Yes	36	423	**92.2%**	21	171	**89.1%**	57	594	**91.2%**
			% Overall	**88.4%**		**% Overall**	**87.5%**		**% Overall**	**88.1%**

Neural Nets Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,456	147	**97.4%**	2,372	60	**97.5%**	7,828	207	**97.4%**
	Yes	286	169	**37.1%**	127	70	**35.5%**	413	239	**36.7%**
			% Overall	**92.9%**		**% Overall**	**92.9%**		**% Overall**	**92.9%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,824	821	**85.5%**	2,059	350	**85.5%**	6,883	1,171	**85.5%**
	Yes	503	5,120	**91.1%**	227	2,169	**90.5%**	730	7,289	**90.9%**
			% Overall	**88.2%**		**% Overall**	**88.0%**		**% Overall**	**88.2%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	375	76	**83.1%**	165	23	**87.8%**	540	99	**84.5%**
	Yes	39	410	**91.3%**	19	187	**90.8%**	58	597	**91.1%**
			% Overall	**87.2%**		**% Overall**	**89.3%**		**% Overall**	**87.9%**

Bayesian Network Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,427	176	**96.9%**	2,353	79	**96.8%**	7,780	255	**96.8%**
	Yes	274	181	**39.8%**	124	73	**37.1%**	398	254	**39.0%**
			% Overall	**92.6%**		**% Overall**	**92.3%**		**% Overall**	**92.5%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,776	821	**85.3%**	2,076	353	**85.5%**	6,852	1,174	**85.4%**
	Yes	524	5,137	**90.7%**	216	2,163	**90.9%**	740	7,300	**90.8%**
			% Overall	**88.1%**		**% Overall**	**88.2%**		**% Overall**	**88.1%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	401	71	**85.0%**	180	30	**85.7%**	581	101	**85.2%**
	Yes	34	426	**92.6%**	13	184	**93.4%**	47	610	**92.8%**
			% Overall	**88.7%**		**% Overall**	**89.4%**		**% Overall**	**88.9%**

LSVM Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,583	20	**99.6%**	2,427	5	**99.8%**	8,010	25	**99.7%**
	Yes	405	50	**11.0%**	180	17	**8.6%**	585	67	**10.3%**
			% Overall	**93.0%**		**% Overall**	**93.0%**		**% Overall**	**93.0%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,823	779	**86.1%**	2,073	340	**85.9%**	6,896	1,119	**86.0%**
	Yes	717	4,944	**87.3%**	308	2,087	**87.1%**	1,025	7,031	**87.3%**
			% Overall	**86.7%**		**% Overall**	**86.5%**		**% Overall**	**86.7%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	389	73	**84.2%**	161	28	**85.2%**	550	101	**84.5%**
	Yes	56	401	**87.7%**	28	174	**86.1%**	84	575	**87.3%**
			% Overall	**86.0%**		**% Overall**	**85.7%**		**% Overall**	**85.9%**

SVM Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,543	60	**98.9%**	2,399	33	**98.6%**	7,942	93	**98.8%**
	Yes	329	126	**27.7%**	137	60	**30.5%**	466	186	**28.5%**
			% Overall	**93.6%**		**% Overall**	**93.5%**		**% Overall**	**93.6%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,919	740	**86.9%**	2,083	325	**86.5%**	7,002	1,065	**86.8%**
	Yes	476	5,135	**91.5%**	206	2,197	**91.4%**	682	7,332	**91.5%**
			% Overall	**89.2%**		**% Overall**	**89.0%**		**% Overall**	**89.1%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	358	50	**87.7%**	197	20	**90.8%**	555	70	**88.8%**
	Yes	52	414	**88.8%**	17	174	**91.1%**	69	588	**89.5%**
			% Overall	**88.3%**		**% Overall**	**90.9%**		**% Overall**	**89.2%**

XGBoost Linear Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,603	0	**100.0%**	2,432	0	**100.0%**	8,035	0	**100.0%**
	Yes	455	0	**0.0%**	197	0	**0.0%**	652	0	**0.0%**
			% Overall	**92.5%**		**% Overall**	**92.5%**		**% Overall**	**92.5%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,658	0	**100.0%**	2,413	0	**100.0%**	8,071	0	**100.0%**
	Yes	5,612	0	**0.0%**	2,389	0	**0.0%**	8,001	0	**0.0%**
			% Overall	**50.2%**		**% Overall**	**50.2%**		**% Overall**	**50.2%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	477	0	**100.0%**	185	0	**100.0%**	662	0	**100.0%**
	Yes	444	0	**0.0%**	208	0	**0.0%**	652	0	**0.0%**
			% Overall	**51.8%**		**% Overall**	**47.1%**		**% Overall**	**50.4%**

Decision List Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	3,844	1,759	**68.6%**	1,671	761	**68.7%**	5,515	2,520	**68.6%**
	Yes	41	414	**91.0%**	14	183	**92.9%**	55	597	**91.6%**
			% Overall	**70.3%**		**% Overall**	**70.5%**		**% Overall**	**70.4%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,438	216	**96.2%**	2,288	98	**95.9%**	7,726	314	**96.1%**
	Yes	2,563	3,055	**54.4%**	1,117	1,304	**53.9%**	3,680	4,359	**54.2%**
			% Overall	**75.3%**		**% Overall**	**74.7%**		**% Overall**	**75.2%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	376	45	**89.3%**	181	23	**88.7%**	557	68	**89.1%**
	Yes	129	330	**71.9%**	58	133	**69.6%**	187	463	**71.2%**
			% Overall	**80.2%**		**% Overall**	**79.5%**		**% Overall**	**80.0%**

C5 Results

		TRAIN			TEST			OVERALL	
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,522	81	**98.6%**	2,391	41	**98.3%**	7,913	122	**98.5%**
	Yes	177	278	**61.1%**	88	109	**55.3%**	265	387	**59.4%**
			% Overall	**95.7%**		**% Overall**	**95.1%**		**% Overall**	**95.5%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,407	245	**95.7%**	2,313	102	**95.8%**	7,720	347	**95.7%**
	Yes	0	5,622	**100.0%**	0	2,386	**100.0%**	0	8,008	**100.0%**
			% Overall	**97.8%**		**% Overall**	**97.9%**		**% Overall**	**97.8%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	386	66	**85.4%**	180	19	**90.5%**	566	85	**86.9%**
	Yes	35	414	**92.2%**	19	183	**90.6%**	54	597	**91.7%**
			% Overall	**88.8%**		**% Overall**	**90.5%**		**% Overall**	**89.3%**

KNN Algorithm Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,570	33	**99.4%**	2,402	30	**98.8%**	7,972	63	**99.2%**
	Yes	242	213	**46.8%**	121	76	**38.6%**	363	289	**44.3%**
			% Overall	**95.5%**		**% Overall**	**94.3%**		**% Overall**	**95.1%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,350	316	**94.4%**	2,299	135	**94.5%**	7,649	451	**94.4%**
	Yes	0	5,610	**100.0%**	0	2,373	**100.0%**	0	7,983	**100.0%**
			% Overall	**97.2%**		**% Overall**	**97.2%**		**% Overall**	**97.2%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	398	66	**85.8%**	162	29	**84.8%**	560	95	**85.5%**
	Yes	30	413	**93.2%**	14	190	**93.1%**	44	603	**93.2%**
			% Overall	**89.4%**		**% Overall**	**89.1%**		**% Overall**	**89.3%**

C&R Tree Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,584	19	**99.7%**	2,429	3	**99.9%**	8,013	22	**99.7%**
	Yes	393	62	**13.6%**	166	31	**15.7%**	559	93	**14.3%**
			% Overall	**93.2%**		**% Overall**	**93.6%**		**% Overall**	**93.3%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,855	786	**86.1%**	2,098	332	**86.3%**	6,953	1,118	**86.1%**
	Yes	577	5,049	**89.7%**	239	2,132	**89.9%**	816	7,181	**89.8%**
			% Overall	**87.9%**		**% Overall**	**88.1%**		**% Overall**	**88.0%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	348	70	**83.3%**	168	27	**86.2%**	516	97	**84.2%**
	Yes	57	403	**87.6%**	19	181	**90.5%**	76	584	**88.5%**
			% Overall	**85.5%**		**% Overall**	**88.4%**		**% Overall**	**86.4%**

Quest Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,585	18	**99.7%**	2,429	3	**99.9%**	8,014	21	**99.7%**
	Yes	393	62	**13.6%**	166	31	**15.7%**	559	93	**14.3%**
			% Overall	**93.2%**		**% Overall**	**93.6%**		**% Overall**	**93.3%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,926	679	**87.9%**	2,161	282	**88.5%**	7,087	961	**88.1%**
	Yes	1,775	3,887	**68.7%**	766	1,592	**67.5%**	2,541	5,479	**68.3%**
			% Overall	**78.2%**		**% Overall**	**78.2%**		**% Overall**	**78.2%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	384	60	**86.5%**	160	23	**87.4%**	544	83	**86.8%**
	Yes	51	400	**88.7%**	26	180	**87.4%**	77	580	**88.3%**
			% Overall	**87.6%**		**% Overall**	**87.4%**		**% Overall**	**87.5%**

Discriminant Analysis Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,561	1,042	**81.4%**	1,959	473	**80.6%**	6,520	1,515	**81.1%**
	Yes	178	277	**60.9%**	88	109	**55.3%**	266	386	**59.2%**
			% Overall	**79.9%**		**% Overall**	**78.7%**		**% Overall**	**79.5%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	4,339	1,301	**76.9%**	1,897	548	**77.6%**	6,236	1,849	**77.1%**
	Yes	2,220	3,424	**60.7%**	975	1,381	**58.6%**	3,195	4,805	**60.1%**
			% Overall	**68.8%**		**% Overall**	**68.3%**		**% Overall**	**68.6%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	346	118	**74.6%**	123	48	**71.9%**	469	166	**73.9%**
	Yes	169	275	**61.9%**	71	131	**64.9%**	240	406	**62.8%**
			% Overall	**68.4%**		**% Overall**	**68.1%**		**% Overall**	**68.3%**

Random Trees Results

		TRAIN			TEST			OVERALL		
No Resampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,222	381	**93.2%**	2,248	184	**92.4%**	7,470	565	**93.0%**
	Yes	41	414	**91.0%**	35	162	**82.2%**	76	576	**88.3%**
			% Overall	**93.0%**		**% Overall**	**91.7%**		**% Overall**	**92.6%**
Oversampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	5,117	520	**90.8%**	2,186	225	**90.7%**	7,303	745	**90.7%**
	Yes	242	5,384	**95.7%**	132	2,258	**94.5%**	374	7,642	**95.3%**
			% Overall	**93.2%**		**% Overall**	**92.6%**		**% Overall**	**93.0%**
Undersampling		No	Yes	**% Correct**	No	Yes	**% Correct**	No	Yes	**% Correct**
	No	407	64	**86.4%**	171	26	**86.8%**	578	90	**86.5%**
	Yes	36	421	**92.1%**	12	196	**94.2%**	48	617	**92.8%**
			% Overall	**89.2%**		**% Overall**	**90.6%**		**% Overall**	**89.6%**