Enhancing Loan Portfolio Management through Multi-Class Classification of Credit Risk: A Case of Kenyan Financial Institutions

By

Crystal Njeri Macharia

078624

Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Crystal Njeri Macharia

23 May 2025

Approval

The dissertation of Crystal Njeri Macharia was reviewed and approved by the following:

Dr. Elphas Okango
Lecturer, Institute of Mathematical Sciences,
Strathmore University

Dr. Godfrey Madigu,
Dean, Institute of Mathematical Sciences,
Strathmore University

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

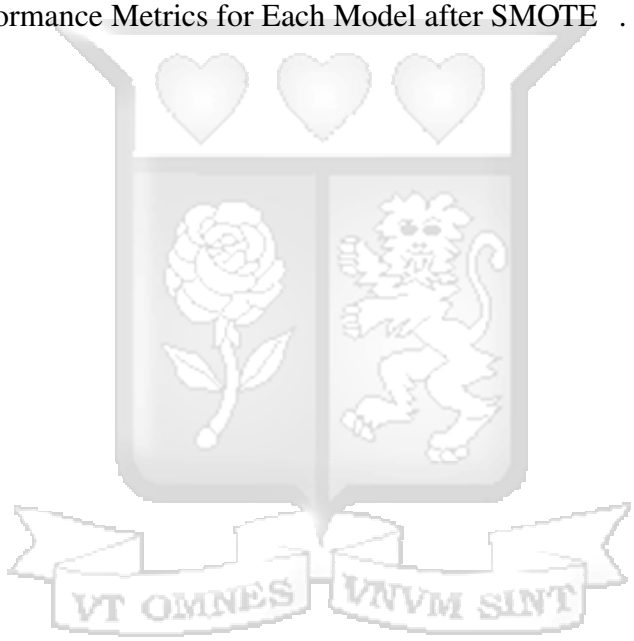
The effective management of loan portfolios and credit risk is crucial for the financial stability of lending institutions. However, recent economic challenges in Kenya have heightened loan default rates, underscoring the need for improved credit risk assessment processes. Traditional methods are increasingly inadequate in addressing evolving market dynamics, prompting some lenders to explore advanced techniques such as machine learning and predictive analytics. Despite their potential benefits, the adoption of these advanced techniques remains limited, particularly among smaller financial institutions. In response to these challenges, this study developed a predictive tool for multi-class loan classification to enhance credit risk assessment and loan portfolio management. Several machine learning algorithms were compared, with XGBoost emerging as the most effective model. The study also evaluated the use of the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance, which improved classification of minority risk categories. The proposed prediction tool aligns with regulatory guidelines and offers a practical solution for lenders to strengthen credit risk monitoring and decision-making, contributing to the resilience and sustainability of financial institutions in Kenya.

Keywords: Credit Risk, Loan Portfolios, Lending Institutions, Loan Default Rates, Machine Learning, Predictive Analytics, Multi-Class Loan Classification, XGBoost, Synthetic Minority Oversampling Technique (SMOTE), Data Imbalance, Decision Making, Kenya

Table of Contents

List of Figures

List of Tables



List of Abbreviations

AI	Artificial Intelligence
AML	Anti-Money Laundering
DL	Deep Learning
DT	Decision Tree
ECL	Expected Credit Loss
EFB	Exclusive Feature Bundling
GBDT	Gradient Boosted Decision Trees
GOSS	Gradient-based One-Side Sampling
KYC	Know Your Customer
LGD	Loss Given Default
LIGHTGBM	Light Gradient Boosting Machine
MFI	Microfinance Institution
ML	Machine Learning
NB	Naive Bayes
PD	Probability of Default
RF	Random Forest
ROA	Return on Assets
SACCO	Savings and Credit Cooperative Organization
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
XGBOOST	Extreme Gradient Boosting

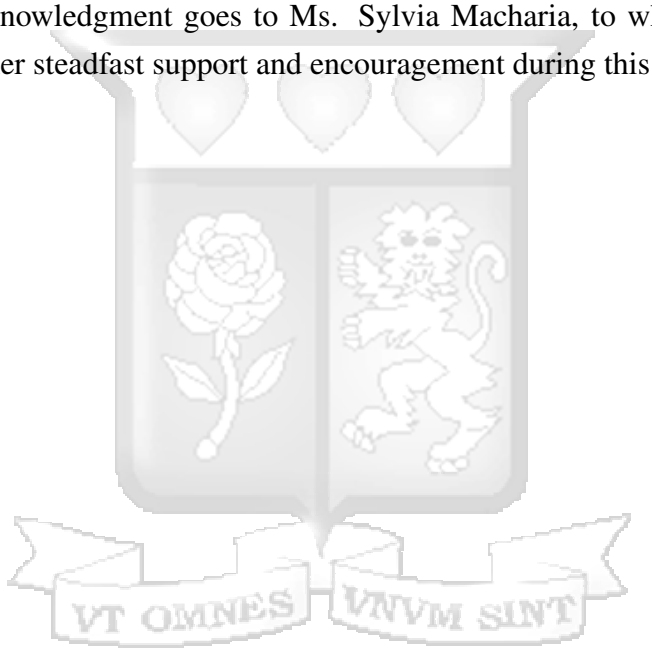
Acknowledgments

I extend my deepest gratitude to God Almighty for His guidance, grace, and favor throughout this journey.

I would also like to sincerely thank my supervisor, Dr. Elphas Okango, for his invaluable guidance, support, and insightful feedback throughout this research. His expertise and encouragement have been instrumental in shaping the direction of this study.

I am profoundly grateful to my classmates, colleagues, friends, and family for their support and motivation throughout this process.

A special acknowledgment goes to Ms. Sylvia Macharia, to whom I am especially indebted for her steadfast support and encouragement during this journey.



Dedication

I dedicate this work to Magdalena Maina, whose unwavering encouragement and belief in me have been a constant source of strength. This achievement would not have been possible without her. I also dedicate it to my son, Teo Ngama, hoping that it inspires him to cultivate a lifelong passion for knowledge and a commitment to excellence.

1. Introduction

1.1 Background to the Study

The effective management of loan portfolios is vital for the financial sustainability and stability, profitability, and growth of lending institutions, particularly in emerging economies like Kenya. Portfolio quality, in particular, plays a key role in driving financial sustainability for microfinance institutions (MFIs). One study found a significant positive relationship between portfolio quality and financial sustainability, suggesting that effective management of loan portfolios is crucial for enhancing profitability and investor confidence in MFIs (Bitok et al., 2020). Loan portfolios represent a significant portion of the assets of a lending financial institution, providing the largest portion of operating revenue (Gongera et al., 2013). These portfolios are subject to various risks including credit risk, liquidity risk, and market risk. Among these risks, credit risk, also known as default risk or performance risk, poses a substantial challenge, as it encompasses the potential for borrowers or counterparties to fail to meet their financial obligations as specified in the agreed terms, leading to financial losses for lenders (Brown and Moles, 2014).

In recent years, the Kenyan economy has faced various challenges, including rising taxes, fluctuating exchange rates, and other economic uncertainties such as high public debt and an elevated cost of living (ReliefWeb, 2023). These challenges have exerted pressure on borrowers' ability to repay loans. Additionally, the implementation of new tax policies and regulatory changes have also contributed to increased financial strain on individuals and businesses, leading to a rise in loan defaults and non-performing assets in the banking sector (Kinyanjui, 2013; Odhiambo, 2013). The COVID-19 pandemic further worsened these challenges, causing disruptions to businesses, loss of employment, and reduced consumer spending, all of which negatively impacted borrowers' repayment capacity. For example, a study on loan repayment behavior revealed that small businesses in Nairobi's Eastleigh area struggled significantly to meet their loan obligations during the pandemic (Noor, 2020). Another study on the impact of mobile loan credit during the COVID-19 pandemic in Kenya revealed that the operations of mobile lending firms were significantly hindered (Oduor, 2021). In response to these economic challenges, lending institutions in Kenya are facing heightened pressure to improve their credit risk assessment processes and mitigate the risk of defaults.

Traditionally, credit risk assessment has relied on a combination of manual processes and analytical tools to assess and evaluate the creditworthiness of counterparties. Credit

analysts typically utilize both financial and non-financial variables, along with various models and analytical techniques, to establish the status of the counterparty (Brown and Moles, 2014). While some methods, such as judgmental approaches and expert systems, involve subjective assessments based on the assessor's experience and lending procedures, others adopt more systematic quantitative techniques, including analytic and statistical models. Additionally, behavioral and market models observe trends over time and rely on financial market prices as indicators of financial solvency. These traditional methods of credit risk assessment are no longer sufficient to address the evolving dynamics of the market. Traditional credit scoring models exhibit limited predictive performance, particularly when applied to large datasets (Bao et al., 2019).

There is, therefore, a growing recognition among lenders and financial institutions of the need to adopt innovative technologies, such as machine learning (ML) and artificial intelligence (AI), to improve the accuracy, efficiency and effectiveness of credit risk analysis. AI models can be leveraged to harness big data for creditworthiness assessments, leading to improvements in predictive accuracy beyond what traditional credit metrics offer (Sadok et al., 2022). Additionally, the evolution of ML algorithms and the accumulation of vast, multidimensional customer data have made the development of credit scoring models using ML techniques a prominent area of interest (Bao et al., 2019). By leveraging advanced data analytics techniques, including predictive modeling, pattern recognition, and data mining, lenders can gain deeper insights into borrowers' creditworthiness, reduce costs associated with credit analysis, proactively identify potential credit risks, take corrective actions to mitigate losses, expedite credit decision-making, and make more informed lending decisions (Ong et al., 2005). These capabilities, in turn, can enhance overall institutional performance.

The current state of the Kenyan economy has led to growing defaults, necessitating closer monitoring of lending portfolios. Despite the potential benefits of adopting advanced credit risk management techniques, the implementation of such strategies remains relatively limited, particularly among smaller financial institutions and microfinance entities in Kenya with challenges such as lack of the appropriate technological infrastructure hindering the widespread adoption of modern credit risk assessment methodologies, leaving many lenders vulnerable to unforeseen credit losses. Addressing these challenges and enhancing loan portfolio management practices are critical imperatives for the Kenyan financial sector. By harnessing the power of data-driven insights and predictive analytics, lending institutions can mitigate credit risks, optimize capital allocation, and ultimately foster profitability. This study seeks to develop an innovative credit risk prediction model to provide accurate and reliable insights into borrower behavior, portfolio management, and potential defaults. Through the utilization of advanced machine learning algorithms and predictive analytics techniques, the

proposed model aims to classify borrowers into one of various risk categories enabling lenders to tailor their risk management strategies as well as their clients' loan portfolios, ultimately empowering lenders with actionable intelligence for making informed lending decisions and effectively managing credit risks in dynamic market conditions.

1.2 Statement of the Problem

In emerging economies like Kenya, the financial stability of lending institutions hinges on their ability to manage loan portfolios and mitigate credit risks effectively. However, the Kenyan economy faces challenges, including a rise in borrower defaults, leading to financial losses for lenders. To address these challenges, lending institutions need to adopt innovative technologies and advanced analytics for credit risk assessment and loan portfolio management. While traditional methods have been useful in the past, they are now insufficient given the complexities of today's market. Despite the potential of machine learning and predictive analytics, their adoption remains limited, especially among small-scale lenders.

Further, guidelines from the Central Bank of Kenya on risk management encourage lenders to develop internal risk rating systems that allow for the effective monitoring and measurement of credit risk, enabling them to make provisions and allocate adequate capital accordingly (CBK, 2013). Yet, the adoption of such systems remains limited, particularly among smaller lenders who lack the necessary infrastructure and resources and therefore lack a structured internal risk rating system altogether or utilize simplistic models that fail to capture the nuances of credit risk adequately. This deficiency hampers their ability to differentiate between various degrees of credit risk and effectively manage their loan portfolios.

Addressing these issues requires a concerted effort to promote the adoption of internal risk rating systems and enhance the quality of management information systems across the lending industry in Kenya, thereby enabling lenders to make informed decisions and proactively manage credit risk. The Central Bank of Kenya suggests the use of a five-category risk rating system including Normal, Watch, Substandard, Doubtful and Loss categories. While the binary classification of credit risk – predicting whether a borrower will default or not – has proven effective in certain contexts, it fails to capture the nuances inherent in borrowers' creditworthiness. Lenders are often caught off-guard when borrowers transition from one risk category to another, moving from being non-defaulters to defaulters overnight without any preemptive signals. This reactive approach to credit risk management underscores the need for a more nuanced classification system, such as a multi-class classification of risks.

By extending the binary solution with more robust risk monitoring and management strategies to encompass a broader spectrum of risk categories such as a multi-class classification framework, lenders can better anticipate and respond to changes in borrowers' credit profiles, mitigating the risk of sudden defaults and enhancing overall loan portfolio management practices. Moreover, there is a critical need for advanced risk analysis techniques to identify shifts in borrowers' risk profiles and facilitate proactive risk monitoring and general planning. A comprehensive risk classification model can provide lenders with the ability to track movement between risk categories over time, providing early warning signs of deteriorating credit quality, enabling lenders to take preemptive measures to mitigate potential losses.

The current study therefore proposes to deploy a model utilizing multi-class classification to address the limitations of traditional credit risk assessment methodologies, with an aim to provide lenders with the ability to proactively manage credit risks and optimize loan portfolio performance.

1.3 Research Objectives

1.3.1 General Objective

The main objective of this paper is to develop and implement an effective credit risk prediction model tailored to the lending landscape in Kenya, by utilizing a multi-class classification technique to enhance loan portfolio management in Kenyan lending institutions.

1.3.2 Specific Objectives

1. To explore data-driven techniques that can be employed for effectively predicting multi-class risk levels in credit assessment.

2. To evaluate multiple machine learning algorithms for multi-class loan classification, assessing their performance in predicting borrower risk categories based on creditworthiness and default probabilities.

3. To investigate the impact of using the Synthetic Minority Oversampling Technique (SMOTE) on an imbalanced dataset in the context of multi-class classification, and to evaluate its effects across various machine learning algorithms.

4. To develop and deploy a predictive tool that utilizes the best-performing model, enabling lenders to accurately classify borrowers into one of multiple risk cate-

gories, improve loan portfolio management, and implement effective risk mitigation strategies.

1.4 Significance of the Study

This study holds significant implications for the financial sector in Kenya and beyond. By developing and deploying a predictive model for credit risk assessment using multi-class classification, this research aims to enhance the way lending institutions manage their loan portfolios and mitigate credit risks. The findings of this study will provide valuable insights and actionable intelligence for lenders, enabling them to make more informed lending decisions, optimize capital allocation, and enhance their overall financial sustainability, stability and profitability. Additionally, the proposed model has the potential to promote financial inclusion by facilitating access to credit for underserved populations and fostering economic development in emerging economies.

1.5 Scope of the Study

This study encompasses a comprehensive examination of credit risk measurement and management practices within the Kenyan financial sector, with a focus on developing and implementing a nuanced, data-driven multi-class classification model. The scope begins with an extensive review of existing literature and best practices in credit risk management, highlighting key challenges and opportunities for improvement. Subsequently, primary data will be collected from a lending institution to inform the development of a predictive model levering machine learning. The analysis and modeling phases focus on testing and evaluating several ML algorithms to evaluate their efficacy in accurately assessing borrower credit risks. Finally, the study extends to the practical application and deployment of the developed model.

1.6 Limitations of the Study

1. The dataset under study pertains to a larger lending institution, which may limit the generalizability of findings to other types of lenders with different operational structures or customer bases.

2. The impact of external factors such as the COVID-19 pandemic and other economic variables may introduce biases or distortions in the data, potentially affecting the accuracy and reliability of the observed lending behavior.

3. Data quality issues, including missing values and redundant columns, could compromise the integrity and completeness of the dataset, potentially leading to inaccuracies or biases in the analysis.

4. Although precautions have been taken to redact or remove sensitive information, the use of customer data raises ethical and privacy concerns, necessitating careful consideration and adherence to relevant regulations and guidelines.

2. Literature Review

2.1 Introduction

Effective loan portfolio management is crucial for sustainable operations and growth in the global financial landscape. Managing credit risk requires precision and adaptability, yet traditional approaches are proving insufficient in today's diverse lending environment. This project aims to enhance loan portfolio management efficiency by implementing advanced machine learning techniques, and building a multi-class classification model that can be utilized in credit risk assessment and classification. By doing so, the project seeks to provide a nuanced approach to understanding customer lending and repaying behavior, and loan portfolio measurement and management. The goal is to offer lending institutions a tool that not only improves risk class prediction accuracy but also enhances decision-making related to loan portfolio management, contributing to the sustainability and success of their lending operations.

2.2 Overview of Loan Portfolio Management

Strong loan portfolio planning and management is a critical function within financial institutions and plays a vital role in the financial performance of lending institutions (Luvuma, 2021). Recognizing the importance of loan portfolio diversification, researchers have shown that effective management of loan portfolios is in direct correlation with reduced credit risk (Aris and Rahimi, 2023). Loan portfolio management refers to the strategic management of a financial institution's portfolio of loans or credits. It involves various processes and activities aimed at optimizing the performance and risk profile of the loan portfolio. A study on loan portfolio management and firm performance defines loan portfolio management as the process of loan portfolio planning, screening of potential borrowers, and credit risk control (Wamalwa and Jagongo, 2017). The study explains that these processes are carried out with the purpose of, among other objectives, achieving high performance and profitability in lending institutions. It also notes that the goal of loan portfolio management is to achieve a well-balanced portfolio that is aligned with the institution's strategic objectives, mitigates potential risks, and ensures financial stability. Another study on the effect of loan portfolio management on financial performance of commercial banks in Kenya found a significant relationship, both positive and negative, between loan portfolio management and the return on assets (ROA) of these banks (Araka, 2022). The collective

works of the authors underscore the vital importance of effective loan portfolio management practices in enhancing financial performance and mitigating risks for lending institutions.

Traditional approaches to loan portfolio management encompass various policies and strategies aimed at optimizing portfolio performance while mitigating risks. One fundamental aspect revolves around assessing the probability of default and loss, tailored to the characteristics and capacities of businesses within specific sub-portfolios (Karekaho, 2009). Another critical policy involves the allocation of various costs, including loan origination, overhead, servicing, and marketing expenses, across different loan types, sizes, and risk levels. Additionally, maximizing shareholder value is emphasized through the creation of risk-efficient portfolios, aiming to balance expected returns with associated risks (Karekaho, 2009).

The implementation of Basel II has prompted a shift in credit risk management, requiring managers to adopt a more quantitative and holistic approach to managing credit portfolios, rather than relying on traditional subjective methods (Vosloo and Styger, 2009). Vosloo and Styger further emphasize the need for an integrated risk management framework that recognizes the interconnectedness of various data points, models, and risk management components. This approach results in proactive risk mitigation, such as anticipating potential issues, and improved credit portfolio analysis.

2.3 Financial Lending in Kenya

Financial lending in Kenya is a dynamic sector that plays a crucial role in driving economic growth and promoting financial inclusion. With a diverse array of financial institutions and lending platforms, including commercial banks, microfinance institutions, and digital lenders, Kenya's lending landscape reflects the pressing need for readily available loans and the availability of diverse lending services across the country, reflecting the dynamic nature of the economy and the widespread demand for financial assistance. The Central Bank of Kenya recognizes the dynamic nature of the financial lending landscape in the country, which has influenced market dynamics, and regulatory frameworks. The institution therefore emphasizes policies for financial stability and consumer protection.

The financial lending sector in the country has experienced substantial growth and diversification with the increase of financial access, specifically in the areas of deposit mobilization, branch expansion, the installation of automated teller machines (ATMs) in rural areas, the introduction of internet banking, and the adoption of electronic systems (Atellu, 2021). The study by Atellu also established that inflation, credit growth

and real interest rate negatively affect financial stability of the country and financial institutions. With the evolving lending landscape in Kenya, there is a growing need for more stringent credit risk assessments in order to ensure lenders' financial stability and profitability. A study on credit risk assessment and loan repayment among development financial institutions emphasize that credit risk assessment is a critical element for ensuring lenders' financial stability and profitability (Okero and Waweru, 2023). The study stresses the importance of accurate evaluation of borrowers' creditworthiness for informed lending decisions, prompting lenders to adapt their assessment strategies to navigate changing economic and regulatory landscapes.

There are several challenges and risks faced by both lenders and borrowers in Kenya's financial lending landscape. One study discusses the challenges facing SACCOs in risk assessment and credit risk management, noting that some SACCOs lack measures to identify relevant issues related to credit risk (Mongina et al., 2022). The study also additionally notes that it is evident that certain SACCOs lack effective credit mitigation techniques to prevent fraud and fund misappropriation. Furthermore, the study notes that high default rates influenced by factors such as economic downturns cause a challenge to lenders. Another study revealed that challenges including high credit facilities' processing fees, strict collateral requirements, and short repayment periods are some of the major hindrances to loans by Kenyans (Gichuki et al., 2014). Additionally, another study raises concerns over the challenge women face in accessing finance despite the existence of over 5,000 registered lending institutions in the country (Makena et al., 2014). In conclusion, both lenders and borrowers in Kenya face significant challenges in the financial lending landscape. These challenges underscore the need for targeted interventions to promote inclusive access to finance and support economic growth and development in Kenya.

2.4 Credit Risk Assessment Methodologies

Traditional credit risk assessment and scoring methodologies have been a key factor in offering a systematic approach to evaluating clients of lending institutions and determining their creditworthiness. These assessments involve a combination of both quantitative and qualitative methodologies where the qualitative methodologies heavily rely on assessor's experience and expert systems such as lending committees. Meanwhile, quantitative methods involve statistical models such as the Z-score method, a method based on the discriminant analysis technique (Brown and Moles, 2014). In their research, Brown and Moles identify various important components of credit risk measurement such as Expected Credit Loss (ECL) - the anticipated loss that a lender or financial institution expects to incur from defaults on loans, Loss Given Default (LGD)

- the proportion of a loan or credit exposure that a lender is expected to lose in the event of a borrower default and Probability of Default (PD) - the likelihood that a borrower will default on a loan obligation within a given time frame. However, these traditional methodologies are facing challenges in keeping pace with the dynamic nature of credit risk. In response to these limitations, there is a growing interest in exploring alternative approaches, such as predictive analytics and machine learning algorithms, to supplement or enhance traditional credit risk assessment techniques.

2.5 Advancements in Credit Risk Assessment - Machine Learning Algorithms for Credit Risk Classification

The conventional methods of credit risk assessment have provided lenders with valuable insights into borrowers' repayment abilities and likelihood of default. However, as the financial landscape evolves and becomes increasingly complex, traditional methodologies are facing challenges in keeping pace with the dynamic nature of credit risk. In response to these limitations, there is a growing interest in exploring alternative approaches, such as machine learning algorithms and predictive analytics. The following section describes the use of such algorithms for classification tasks, including credit risk classification.

2.5.1 Decision Trees

A decision tree is a classifier structured like a flowchart, where internal nodes represent tests on input attributes and branches signify the outcomes of those tests (Rokach and Maimon, 2005). The leaves of the tree correspond to class labels or probability vectors for the target attribute. This hierarchical and recursive structure allows the instance space to be partitioned into subspaces based on attribute values (Rokach and Maimon, 2005). Another definition describes Another definition describes decision trees as classifiers that predict class labels by posing a series of questions about the features of data items. These questions form a hierarchy encoded as a tree, where each node represents a question and each branch represents a possible answer, an approach valued for its simplicity and interpretability (Kingsford and Salzberg, 2008). A paper titled "Early Prediction of Hypothyroidism and Multiclass Classification Using Predictive Machine Learning and Deep Learning" investigates the effectiveness of machine learning (ML) and deep learning (DL) methods in predicting and classifying hypothyroidism. The authors point out the rising incidence of thyroid disorders, especially in India, and stress the importance of early and precise diagnosis to minimize health risks. They assess the performance of various models in distinguishing between dif-

ferent types of hypothyroidism (negative, compensated, primary, and secondary) and find that the Decision Tree model yields the highest accuracy of 99.5758% concluding that Decision Trees show considerable potential for the accurate and timely detection of hypothyroidism (Guleria et al., 2022). Another paper titled "Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study," decision trees are employed as a classification algorithm to predict loan defaults within the banking sector. The researchers trained a decision tree model on a dataset of approximately 2.2 million loans, achieving an accuracy of 73%, concluding that the algorithm provide valuable insights in the realm of credit risk and loan default prediction (Madaan et al., 2021).

2.5.2 Random Forests

Random Forests (RF) is an ensemble machine learning technique that has been widely leveraged as a tool to classify credit scores and to establish the risk of default. The algorithm involves the composition of multiple decision trees, each trained on a random subset of the data. This introduces randomness into the tree-building process, resulting in a diverse set of trees that collectively make predictions by voting for the most popular class. As a result, random forests often achieve significant improvements in classification accuracy compared to individual decision trees (Breiman, 2001). In a study comparing RF with other statistical classifiers on invasive plant species data, RF exhibited consistently high classification accuracy. The study highlights the various advantages of the algorithm such as superior accuracy, robust handling of complex variable interactions, and versatility in performing various types of statistical analysis Cutler et al. (2007). RF has also been extensively employed for multi-class classification tasks. For instance, one study used RF to classify higher education students into multiple categories based on their first-year results (Berriri et al., 2021). Similarly, the technique has been applied to the multi-class classification of groundnut disease (Chaudhary et al., 2016).

2.5.3 Gradient Boosted Decision Trees

Gradient Boosted Decision Trees (GBDT) stand out as widely favored supervised learning techniques in various industries (Saberian et al., 2019). Not only do they boast high accuracy, but they also demonstrate rapid prediction capabilities, interpretability, and a minimal memory footprint. GBDT sequentially build a series of decision trees, with each tree aiming to rectify the errors of its predecessor. Starting with a single decision tree, additional trees are progressively fitted. The ensemble's final prediction is derived by aggregating the predictions of all the trees in the sequence (Friedman,

2001). GBDT have been widely applied in credit risk classification, emerging as the most efficient model for this task, outperforming other methodologies (Tian et al., 2020). Additionally, through the utilization of the Grid Search algorithm for parameter adjustment, the algorithm achieved improved solutions in credit quality classification (Tian et al., 2020). Another study notes that tree-ensembled frameworks offer a balanced approach concerning performance, efficiency, and interpretability in credit scoring (Liu et al., 2022). The study notes that, GBDT notably enhances credit scoring performance by iteratively optimizing credit scoring errors. GBDT has similarly been explored for various multi-class classification tasks. For instance, the algorithm has been employed the algorithm for multi-class classification of tweets for sentiment analysis and opinion mining, successfully categorizing the data as negative, positive, or neutral (Neelakandan and Paulraj, 2020). Similarly, it has been leveraged for one-month-ahead prediction of drought conditions, categorized into three classes: wet, normal, and dry events (Danandeh Mehr, 2021).

2.5.4 Extreme Gradient Boosting (XGBoost)

XGBoost, short for eXtreme Gradient Boosting, is a scalable and efficient tree boosting system that implements the gradient boosting framework, a machine learning technique that combines multiple weak prediction models, typically decision trees, to create a stronger, more accurate model (Friedman, 2001). The XGBoost algorithm works by first making an initial prediction, often a simple average, to establish a baseline. It then computes the gradient of the loss function, which indicates how to adjust the model to reduce errors. A new decision tree is then constructed to predict these gradients, aiming to identify patterns in the model's mistakes. This tree is added to the existing ensemble with a weight determined through optimization, and the overall model is updated. This process continues iteratively until performance improvements plateau or a set number of iterations is reached (Chen et al., 2015).

In a study on Credit Debt Default Risk Assessment Based on the XGBoost Algorithm, the effectiveness of XGBoost is realized in predicting credit bond defaults (Wang et al., 2022). The researchers analyzed a dataset of credit bond issuances from 2014 to 2020, with the aim of creating a model to accurately identify those likely to default. They leveraged the XGBoost model and noted significant improvements in predictive accuracy of the optimized version over the baseline model, demonstrating its potential as a valuable tool for assessing credit debt default risk in China (Wang et al., 2022). In another study titled 'XGBoost for Imbalanced Multiclass Classification-Based Industrial Internet of Things Intrusion Detection Systems', the authors investigate the application of the XGBoost algorithm for detecting intrusions in Industrial Internet of

Things (IIoT) networks, focusing on the issue of imbalanced multiclass datasets (Le et al., 2022). They point out the shortcomings of deep learning models, such as RNNs and LSTMs, which often perform poorly in detecting attacks when the distribution of attack types is uneven, or imbalanced, tending to favor majority classes and neglecting minority ones. In contrast, the XGBoost model showcases impressive accuracy and efficiency, effectively tackling these challenges and emerging as a valuable solution for enhancing intrusion detection in IIoT systems (Le et al., 2022).

2.5.5 Light Gradient Boosting Machine (LightGBM)

In recent years, GBDT has encountered challenges, particularly regarding the balance between accuracy and efficiency. Conventional implementations of GBDT exhibit substantial time consumption when confronted with large data sets. In response to this problem, the Light Gradient Boosting Machine (LightGBM) algorithm emerged as a strategic solution to expedite the training process by up to over 20 times while preserving nearly the same accuracy (Ke et al., 2017). LightGBM is an efficient and high performance gradient boosting framework developed by Microsoft that employs decision tree algorithms to construct ensemble models. Its primary aim is to increase processing speed, enhance model accuracy, and to minimize memory usage. LightGBM has emerged as a powerful tool in the domain of credit risk assessment, as evidenced by its successful application in several research studies. One study focused on borrower credit classification within P2P network loans, highlighting LightGBM's superiority in achieving high accuracy in data classification. The simulation experiments demonstrated the model's stability, fitting ability, and remarkable accuracy in classification prediction (Zhang et al., 2019). Another study aimed to construct a financial default risk prediction model for mitigating risks based on the algorithm. The results show that the algorithm achieved an accuracy rate of 80.25% on the test set, outperforming models like logistic regression and support vector machines (Gao and Balyan, 2022). Additionally, a comparative study of popular algorithms underscored LightGBM's efficacy in credit risk analysis, displaying better prediction and higher stability (Ponsam et al., 2021). These findings collectively underscore LightGBM's prowess in credit risk assessment tasks and its potential to revolutionize risk management practices in the financial sector.

2.5.6 Support Vector Machines

Support Vector Machines (SVM) have become a popular machine learning algorithm due to their relative simplicity and flexibility in handling classification tasks (Pisner and Schnyer, 2020). SVM works by finding the optimal separating hyperplane that

best separates different classes in the feature space by maximizing the margin (Meyer and Wien, 2001). The algorithm has been employed widely for both binary and multi-class classification. For instance, one study utilized SVM for arrhythmia detection using digital images of cardiac signals and R-R intervals, achieving an accuracy of 81.9% across four classes: cardiac normal, atrial premature beat arrhythmia, atrial flutter arrhythmia, and atrial fibrillation arrhythmia (Hangkawidjaja et al., 2021). Another study employed SVM for detecting bearing defect in rotating machinery, demonstrating the effectiveness of a developed sensor in identifying different classes of bearing damage. SVM results validated the sensor as a cost-effective tool for condition monitoring (Goyal et al., 2020). Additionally, another study applied SVM for lung nodule segmentation in cancer identification systems, emphasizing its role in accurate classification of lung cancer staging based on morphological variations of lung nodules, thereby enhancing disease screening accuracy (Lavanya et al., 2021).

2.5.7 Naive Bayes

Naive Bayes (NB) classifiers, a fundamental component of machine learning, are based on the Bayes theorem with the ”naive” assumption of independence among features. This assumption simplifies the computation of probabilities and allows for efficient classification. Despite its simplicity, NB is very computationally efficient and delivers high classification accuracy (Webb et al., 2010). NB classifiers have been extensively utilized across various domains, with notable applications as highlighted by Wickramasinghe and Kalutarage (2021). In software defect prediction, NB emerges as the most prevalent learner group, constituting 47.4% of all studies, showcasing its widespread adoption and effectiveness in this field. Moreover, NB has been employed in decision support systems for predicting heart disease, demonstrating its utility in healthcare applications. Additionally, NB classifiers play a significant role in predicting liver diseases, further underscoring their relevance and impact in health-related fields. In the realm of education, NB algorithms have been instrumental in developing web-based systems for forecasting students' academic success. These systems enable educators to identify students at risk of unsatisfactory performance, allowing for timely intervention and support (Wickramasinghe and Kalutarage, 2021). Various studies have demonstrated the versatility of NB in credit scoring applications. A computational study by focusing on constructing a predictive credit scoring model using publicly available German credit data showcased NB's efficacy in this domain (Trivedi, 2020). Additionally, NB was successfully employed in the detection of credit card fraud, highlighting its adaptability in addressing financial security challenges (Itoo et al., 2021). In a comprehensive survey paper, a systematic review of existing research methodologies and machine learning techniques for credit risk evaluation was

conducted. The findings underscored NB as a popular and effective model utilized across a spectrum of credit scoring tasks (Bhatore et al., 2020).

2.5.8 Multiclass Classification

Multiclass classification is a type of supervised machine learning where instances are categorized into one of three or more discrete classes (Del Moral et al., 2022). Unlike binary classification, which distinguishes between only two possible outcomes (e.g., default vs. no default), multiclass classification is capable of assigning each instance to a single category among multiple possible outcomes. This approach is widely used in areas such as image recognition, sentiment analysis, and medical diagnosis, where distinguishing between more than two outcomes is essential for accurate and interpretable decision-making.

In the context of borrower risk monitoring, loan portfolio management, and credit risk classification, multiclass classification offers significant advantages. While binary models classify borrowers into broad categories such as 'defaulter' or 'non-defaulter', multiclass models enable more granular categorization of borrowers, allowing for the assignment of borrowers into intermediate risk categories based on their credit profiles. This added granularity deepens credit risk analysis and supports financial institutions in implementing more tailored and proactive lending strategies.

Additionally, the technique enhances the accuracy of predicting not just the likelihood of default but also the likelihood of a borrower transitioning between different risk states over time. This approach allows for the continuous monitoring of borrowers' financial health, helping to anticipate changes in their creditworthiness. One study demonstrated the use of a deep learning-based multiclass model to classify borrowers into three distinct risk groups: high risk (charged-off or defaulted), medium risk (late-paying), and low risk (fully paid). The study found that this approach not only improved prediction accuracy but also enabled better identification of borrowers likely to transition between risk states (Paudel et al., 2023a).

Further supporting this view, another study proposed an ensemble learning model based on multi-source fusion theory that categorizes borrowers into five credit risk levels. Their work underscores the idea that creditworthiness is rarely binary; rather, borrowers often fall into nuanced categories requiring differentiated risk treatment. This reinforces the practical value of multiclass classification for more effective segmentation and credit decision-making(Wang et al., 2024).

In conclusion, multiclass classification offers a powerful tool for improving loan portfolio management and credit risk classification. By providing more granular insights

into borrower behavior and credit risk, multiclass models enable financial institutions to allocate resources more efficiently, reduce risks, and optimize their overall lending strategies. Moreover, multiclass classification supports regulatory requirements that demand more differentiated risk assessments and aligns well with the growing use of machine learning in financial services (CBK, 2013). The ability to capture varying levels of borrower risk not only improves model interpretability and decision-making but also supports long-term financial stability and compliance objectives.

2.6 Synthetic Minority Oversampling Technique (SMOTE)

In many real-world applications, there is a prevalence of imbalanced datasets. Imbalanced datasets are those in which the classification categories are not approximately equally represented, meaning that one class, referred to as the majority class, has significantly more samples than another class, known as the minority class (Elreedy et al., 2024; Wei et al., 2022). The issue with imbalanced datasets is that traditional classification algorithms tend to focus more on majority class samples, often overlooking minority class samples. Despite their smaller numbers, minority class samples can contain valuable information (Wei et al., 2022). Additionally, traditional performance metrics like accuracy can be misleading in this context; a classifier can attain high accuracy by predominantly predicting the majority class, without effectively identifying minority class instances (Chawla et al., 2002).

Researchers have proposed and developed various methods to address these limitations, including under-sampling and over-sampling with replacement. Under-sampling involves randomly removing samples from the majority class to achieve a more balanced class distribution, while over-sampling with replacement entails duplicating existing minority class samples to enhance their representation (Elreedy and Atiya, 2019). The drawback of these methods is that under-sampling can result in the loss of important information, while over-sampling poses the challenge of not knowing the underlying distribution. This uncertainty makes it difficult to accurately emulate the distribution when generating new data, thus making over-sampling with replacement a more intuitive approach (Elreedy and Atiya, 2019). However, Chwala argues that over-sampling with replacement does not significantly improve recognition of the minority class (Chawla et al., 2002). This technique merely creates more specific decision regions around minority samples without expanding the overall decision boundary, which can lead to overfitting, a scenario where the classifier performs well on training data but poorly on unseen data.

To overcome these limitations, a more effective over-sampling technique, Synthetic Minority Over-sampling Technique (SMOTE), was developed (Chawla et al., 2002). The over-sampling method has become extremely popular and successful for generating new data. SMOTE works by creating synthetic minority samples through interpolation between existing minority samples and their nearest neighbors in feature space, effectively increasing the representation of the minority class without merely duplicating existing instances. This technique addresses the challenges of imbalanced datasets by encouraging the creation of more general decision boundaries, thus leading to improved performance of classifiers.

Despite the efficacy of SMOTE, the over-sampling algorithm has some limitations. One limitation of SMOTE is that it can amplify noise in the dataset by oversampling noisy examples, thus negatively impacting classification performance, as many oversampling techniques struggle to perform effectively in the presence of noise (Wei et al., 2022). Researchers advocate for filtering noise during the data preprocessing stage before applying oversampling techniques. For example, one study proposed the IR-SMOTE technique which uses K-means clustering and a distance metric to identify and remove noise samples within the minority class (Wei et al., 2022). Another limitation of SMOTE is the lack of a solid mathematical foundation for the algorithm, which makes it difficult to analyze the relevance and distributional properties of the synthetically generated samples (Elreedy et al., 2024). Researchers are focused on deriving a mathematical framework for the probability distribution of SMOTE-generated synthetic samples to better understand their relevance and proximity to the true data distribution.

Several variations of SMOTE have been developed to address its limitations such as noise amplification and overgeneralization. Borderline-SMOTE, improves SMOTE by focusing on only the minority class instances that lie near the decision boundary, as these are more likely to be misclassified (Han et al., 2005). The technique therefore prioritizes those in ambiguous regions, reinforcing separation of classes. Another approach, ADASYN (Adaptive Synthetic Sampling), enhances SMOTE by generating synthetic samples adaptively, assigning more new instances to minority samples that are harder to classify, by shifting the decision boundary towards the more difficult minority samples (He et al., 2008). SMOTE-NC (Nominal-Continuous), modifies the interpolation mechanism to handle datasets containing both numerical (continuous) and categorical (nominal) features. Instead of applying linear interpolation to categorical attributes, it selects new values based on the most frequent category among nearest neighbors, making it more suitable for categorical variables and mixed-type datasets developed by (Chawla et al., 2002),. These refinements to SMOTE show the ongoing efforts to improve the oversampling technique, making it more resistant to noise,

adaptive, and suitable for complex datasets. While these variations introduce improvements, standard SMOTE remains a widely adopted baseline due to its simplicity and effectiveness.

2.7 Ethical, Regulatory and Compliance Issues

There is growing recognition of key regulatory and ethical considerations in the application of machine learning in finance. One study stresses the need to address data privacy and security concerns, recognizing the sensitivity of customer data handled by machine learning models (Aris and Rahimi, 2023). The study emphasizes that compliance with data protection laws as paramount for financial institutions, underscoring the importance of safeguarding customer information and preventing unauthorized access to uphold regulatory standards. Explainability and transparency in lending decisions have also been identified as critical, particularly given the complexity of machine learning models. To meet regulatory expectations, including providing clear explanations for decisions, institutions must be able to provide clear justifications for decisions, reducing perceptions of models as opaque "black boxes" and ensuring fair treatment of borrowers (Misheva et al., 2021). Another study highlights the necessity for model validation and governance, noting potential regulatory mandates for robust frameworks to assess the performance, accuracy, and fairness of machine learning models (Umagba et al., 2022). This requirement ensures that lending institutions can demonstrate the reliability of their models while adhering to established regulatory standards. The integration of machine learning with regulatory obligations such as Anti-Money Laundering (AML) and Know Your Customer (KYC) requirements has also been explored. While these models enhance AML and KYC processes, alignment with compliance standards remains essential to protect the integrity of customer data and prevent illicit activities (Rehman et al., 2019).

As the use of data continues to grow across various sectors, ethical concerns regarding data collection, usage, and privacy have become more pronounced. One major concern is informed consent, where individuals may not fully understand how their data is being used, potentially infringing on their autonomy and privacy. Scholars have argued that conventional 'notice and consent' frameworks often fail to offer genuine protection, as users may agree to terms without fully comprehending their implications (Nissenbaum, 2010). Ethical guidelines stress the importance of transparent data collection practices and ensuring that individuals maintain control over their personal information. Additionally, issues such as bias in data mining and decision-making algorithms have raised concerns about fairness and discrimination. Researchers highlight the potential for machine learning models to perpetuate existing biases in training data, leading to

discriminatory outcomes (Barocas and Selbst, 2016). The concept of accountability plays a critical role in data ethics, particularly as algorithms are increasingly used to make decisions that significantly affect individuals' lives. There have been growing calls for institutions to be held responsible for these decisions, ensuring that algorithmic outcomes are not only transparent but also fair and explainable. It is essential for institutions to take responsibility for the impacts of algorithmic decision-making, especially given the potential for algorithms to perpetuate biases and discrimination if not properly managed (Shah, 2018). Data privacy is another key issue, especially with the increasing use of personal data by corporations and governments. Ethical frameworks emphasize the importance of protecting individuals' right to privacy, ensuring that data is not misused, and enforcing strict data protection measures to prevent unauthorized access and breaches. However, as technology advances and more personal data is collected, there is a need to recalibrate our relationship with privacy. The unprecedented possibilities unlocked by vast amounts of personal data necessitate a new understanding of privacy, one that not only protects us against harm, but also upholds the fundamental principle of respect for persons in the digital age (Seynhaeve, 2022). As data continues to play a central role in decision-making processes, addressing these ethical challenges is crucial for fostering trust and ensuring that data-driven solutions are both fair and responsible.

2.8 Gaps in Existing Techniques

Traditional credit risk assessment methods and binary classification models, while effective in some scenarios, fall short in capturing the complex and dynamic nature of borrower behavior. These limitations highlight the need for more nuanced and adaptive approaches, such as multiclass classification.

Traditional credit risk metrics like Expected Credit Loss (ECL), Loss Given Default (LGD), and Probability of Default (PD) present some challenges. While these metrics are foundational in risk modeling, they often rely on static assumptions that may not adequately reflect the fluid nature of economic conditions or individual borrower circumstances. This can limit their effectiveness in real-time credit risk monitoring. In contrast, machine learning techniques such as multiclass classification can dynamically adapt to changes in borrower data, offering more responsive and insightful predictions.

Binary classification methods, which categorize borrowers into only two groups, are often too simplistic to reflect the complexity of real-world borrower behavior. One paper which explores and illustrates the use of deep learning models in multi-class evaluation problems recognizes that a simple binary classification lacks the necessary detail for a more nuanced understanding of credit risk (Paudel et al., 2023b). Addition-

ally, binary models fail to capture transitions between risk categories and often results in reactive risk management. Lenders may be caught off guard when a borrower transitions rapidly from a stable to a defaulting state without prior indication. In contrast, multiclass classification offers the ability to detect and differentiate between multiple risk levels, enabling proactive intervention and tailored credit strategies.

Another significant issue arises from the imbalance commonly found in credit datasets, where non-default cases vastly outnumber default cases. This imbalance can skew model predictions and reduce performance. While techniques like under-sampling and over-sampling have been used to address this, they each have drawbacks: under-sampling risks discarding valuable information, whereas over-sampling can introduce redundancy or noise. These challenges make it difficult to accurately emulate the distribution when generating new data (Elreedy and Atiya, 2019). The development of more sophisticated methods such as the Synthetic Minority Over-sampling Technique (SMOTE) offers a more balanced approach by creating synthetic examples of minority classes, improving the learning process without compromising data quality.

Together, these limitations underscore the importance of evolving beyond conventional methods. Multiclass classification, especially when integrated with modern sampling techniques and machine learning frameworks, offers a more robust strategy for managing credit risk in today's dynamic lending environments.

2.9 Summary

The reviewed literature provides valuable insights related to the application of machine learning in loan portfolio management in Kenya. It offers a comprehensive exploration of loan portfolio management, the financial landscape in Kenya, credit risk assessment methodologies, advancements in credit scoring, and regulatory and compliance concerns. The literature review showcases the pivotal role of loans for financial institutions in generating income, emphasizing the importance of loan portfolio management for risk management and financial growth and stability. The literature highlights that while traditional methods of risk measurement are widely accepted and utilized in lending institutions, machine learning methods are more effective in predicting the accuracy of possible credit risk and for classifying borrowers.

3. Methodology

3.1 Introduction

The methodology section serves to outline the processes undertaken to achieve the objectives of this research. It provides a detailed explanation of the research methods, data employed and algorithms utilized. This aims to ensure that the reader comprehends the process and procedures involved in conducting the analysis, thereby validating the reliability and replicability of the findings. The research design adopts a quantitative approach to develop and evaluate a credit risk prediction model, leveraging data sourced from a prominent financial institution in Kenya while prioritizing privacy and regulatory compliance. Rigorous data preprocessing techniques, including cleaning and feature selection, are employed to optimize model performance. The study utilizes a diverse set of machine learning algorithms including, Decision Trees, Random Forests, and Gradient Boosted Decision Trees. Evaluation metrics encompass the confusion matrix and cross-validation techniques, ensuring comprehensive assessment of model effectiveness. This approach aligns with the Cross Industry Standard Process for Data Mining (CRISP-DM), a recognized framework guiding data mining projects through structured phases of exploration, modeling, and deployment (Wirth and Hipp, 2000). CRISP-DM encompasses several key phases as seen in fig. 3.1, including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. In the Business Understanding phase, project objectives, requirements and success criteria are defined. Data Understanding involves exploring and familiarizing oneself with the available data. Data Preparation focuses on cleaning, transforming, and formatting the data for analysis. The modeling phase involves selecting and applying appropriate modeling techniques. Evaluation assesses the effectiveness and performance of the models. Deployment involves integrating the models into operational systems and processes. This iterative process allows for thorough exploration, modeling, and validation of data mining solutions.

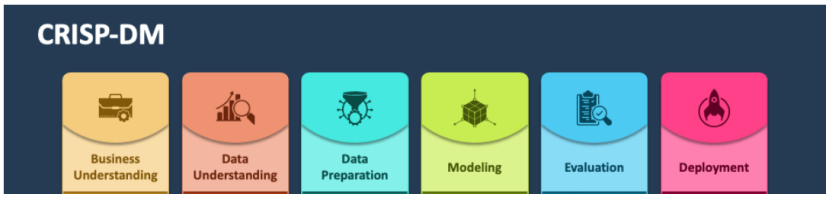


Figure 3.1: CRISP-DM Framework
Source: (Oluwagbenga, 2023)

3.2 Business Understanding

For lending institutions, accurate credit risk assessments and loan status classifications is critical for their sustainability and stability, profitability, and growth. The goal of this project is to develop a robust machine learning model that can effectively predict loan classifications, helping financial institutions make informed decisions about credit risk. By leveraging advanced multi-class classification and oversampling techniques, this project aims to enhance the accuracy and effectiveness of loan portfolio management within lending financial institutions. This project focuses on evaluating various machine learning models to identify the most effective approach for multi-class loan classification. Additionally, it aims to enhance predictive accuracy by leveraging oversampling techniques to ensure adequate representation of minority classes. The requirements of this project include a well-structured dataset with labeled loan classifications as well as appropriate tools and suitable techniques and methodologies for efficient experimentation and analysis. The success of this project is determined by its ability to develop a reliable loan classification model while effectively handling class imbalances. The model should achieve better performances than a baseline trained on imbalanced data, with notable improvements in minority class predictions. It must generalize well across all loan categories. The solution should be interpretable and practical, providing meaningful insights and reliable predictions that enable financial institutions to perform more accurate loan classification and risk assessment.

3.3 Data Understanding

The data utilized in this project was obtained from one of the leading commercial banks in Kenya, a prominent financial institution that has requested to remain anonymous. The dataset was retrieved in the form of a flat file, specifically an XLSX file, and contains comprehensive information on various financial transactions, accounts, and credit profiles. It provides extensive details on financial attributes related to the institution's lending operations, including loan product type, balances and exposures, overdraft days, and risk classification category. The dataset covers a wide range of over 19,000 loans disbursed to the institution's clients. While the dataset is authentic and based on real financial records, the specific internal processes used by the bank to collect and compile the data were not disclosed. To ensure data privacy and comply with confidentiality regulations, all sensitive client information, including personal identifiers, was redacted or removed prior to data sharing. The dataset encompasses various features used in this analysis as listed in table 3.1:

Table 3.1: Data Description

Column/Variable	Description
Product Line	This column categorizes the type of loan product taken by the borrower. It includes various loan types such as mortgages, asset finance, commercial term, business, and digital. Additionally, it may indicate whether the loan is secured or unsecured, providing insights into the collateral requirements for each loan type.
Business Unit	This column indicates the division or unit within the financial institution that handles the loan. It includes the categories business banking, corporate banking, and personal banking.
Segment	This column represents the sub-categories or specialized divisions within each broader business unit, delineating specific areas of focus or customer segments.
Disbursed	This column represents the total amount of funds lent or granted to the borrower at the time of loan disbursement. It provides information on the initial loan amount extended to the borrower.
Arrears	This column denotes the cumulative amount of overdue payments or outstanding debt owed by borrowers at a given point in time. It reflects the aggregate sum of missed or late payments beyond the scheduled due dates.
Exposure	The exposure column reflects the total financial exposure or liability associated with the loan. It may include the outstanding balance, accrued interest, and any other associated fees or charges, providing a comprehensive view of the borrower's financial obligation to the institution.
OD Days (Overdraft Days)	This column records the number of days the borrower has exceeded the agreed overdraft limit. It measures the duration of time during which the borrower has utilized overdraft facilities beyond the approved limit.
Risk Class Held	This column categorizes the credit risk associated with each loan based on predefined risk classes. These risk classes include "doubtful," "loss," "watch," "normal," and "substandard," each representing varying levels of creditworthiness and likelihood of default. This column serves as a key variable for assessing and managing credit risk within the loan portfolio.

3.4 Data Preparation

3.4.1 Data Cleaning

The data cleaning pipeline was guided by the revelations brought about by performing preliminary exploratory data analysis (EDA) tasks, which serve as a tool to gain a better understanding of the data. The preliminary EDA tasks included reviewing the number of rows and columns in the data, understanding the balance or lack thereof of the classes in the target variable and viewing the distributions and summary statistics of the data. Data cleaning tasks performed were as follows:

1. Handling missing values: Missing values refer to the absence of data in certain

columns or rows within a dataset. Dealing with missing values is crucial as they can affect the accuracy and reliability of analytical results. The Python Programming Language library pandas function `.isnull()` was used to check for null values. For handling missing data, imputation was applied where possible. In cases where missing values were minimal and a logical replacement was available, imputation was preferred over deletion to preserve as much data as possible. However, for variables with a large proportion of missing values (more than 50% missing values), dropping the columns was the most appropriate approach to avoid introducing bias through imputation.

2. Treating duplicate data: Duplicate data occurs when there are identical records present in a dataset, which can skew analytical results. Pandas library offers a function `.duplicated()`, which was used to check for duplicate values. Since duplicates can lead to misleading analysis and redundancy, all exact duplicate rows were dropped using the `.drop_duplicates()` function. This helped maintain a clean and accurate dataset for further processing.

3. Correcting data types: Data types define the nature of values stored in each column of a dataset, such as integers, floats, strings, or datetime objects. Ensuring that data types are correctly assigned is critical for performing accurate calculations, operations, and analyses. Using pandas' `.dtypes` attribute, the data types of columns were inspected to identify any inconsistencies or mismatches. To address these inconsistencies, necessary conversions were performed using the function `.astype()` for both the numerical and categorical variables that needed adjustments ensuring consistency in the data.

4. Dropping Columns: Some columns in a dataset may contain irrelevant, redundant, or unnecessary information that does not contribute to the analysis or modeling process. Dropping such columns helps streamline the dataset, reduce dimensionality, and improve computational efficiency. The pandas' `.drop()` function was employed to remove columns from the dataset based on their relevance to the study, or their high correlation with other variables, which could introduce redundancy.

3.4.2 Exploratory Data Analysis

With the dataset now cleaned and prepared, the analysis progresses to the phase of exploratory data analysis (EDA). This critical step involves delving deeper into the dataset's characteristics and patterns to uncover insights that may inform subsequent analytical decisions. Through visualizations, statistical summaries, and in-depth ex-

amination, the EDA aims to reveal underlying relationships and trends within the data, laying the foundation for more advanced modeling and interpretation

A. Univariate Data Analysis

A comprehensive univariate exploratory data analysis (EDA) was conducted to gain insights into the individual features of the dataset. Univariate analysis involves examining each variable in isolation, allowing for the exploration of its distribution, central tendencies, and variability. This approach is crucial for gaining a foundational understanding of the data's structure and uncovering valuable patterns and trends. The seaborn and matplotlib libraries from the Python programming language were leveraged to plot various charts for this EDA. For categorical columns, the univariate analysis primarily focused on assessing the distribution of each category within the dataset. Bar plots, pie charts and count plots were utilized to visualize the frequency of occurrences for each category, ensuring that class imbalances or rare categories were identified early in the process. Additionally, for numerical columns, summary statistics such as mean, median, standard deviation, minimum, and maximum values were computed to provide a comprehensive overview of their central tendencies and variability. Identifying skewed distributions or extreme values helped determine whether transformations (e.g., normalization) were necessary before applying machine learning algorithms.

B. Bivariate Data Analysis

The bivariate analysis primarily focused on exploring relationships between the target variable, `Risk Class Held`, and other variables in the dataset. The goal of this step was to examine how the target variable interacts with various predictors, such as loan characteristics, identifying those that contribute to risk classification. Various plots and charts were used to understand these relationships such as clustered bar charts which were used to observe how categorical features varied across different risk classes, helping to identify strong categorical predictors. The insights gained at this stage guided further preprocessing steps, such as encoding categorical variables, ensuring optimal model performance.

C. Multivariate Data Analysis

Multivariate analysis involves examining the relationships between multiple variables simultaneously. It allows for a comprehensive understanding of how various factors interact and influence each other within a dataset. In this study, multivariate analysis was conducted to uncover complex patterns and dependencies among the predictors, target variable, and other relevant variables. The visualizations done include a pairplot, which displays pairwise relationships among the

selected numerical variables, and a correlation heatmap, a visual representation of the correlation matrix, showing the strength and direction of linear relationships between all pairs of variables. By performing this multivariate analysis, the dataset was refined to include only the most relevant predictors, ensuring that the final model was both interpretable and efficient.

3.5 Machine Learning Modeling

3.5.1 Data Preprocessing

Data preprocessing steps are essential in preparing the dataset for modeling, ensuring that the data is in a suitable format and condition for machine learning algorithms to effectively learn from it, and make accurate predictions. Various preprocessing steps were employed in this research as follows:

a) Dropping high cardinality columns: High cardinality in categorical variables can pose challenges for machine learning algorithms, particularly when using techniques like one-hot encoding, which would result in an excessively large dataset. Additionally, the sheer number of unique values in these columns could introduce noise and complexity into the model without necessarily adding significant predictive power.

b) One-hot encoding: The technique is used to convert categorical variables into a numerical format suitable for machine learning algorithms. It works by creating binary columns for each category in the original variable, where a '1' indicates the presence of the category and '0' indicates absence. To perform one-hot encoding, the `.get_dummies()` function is used.

c) Separating the dataset: This segregation facilitates the training and evaluation of machine learning models. Features (**X**) comprise all columns except the target variable, while the target variable (**y**) is isolated for classification purposes.

d) Scaling the numerical variables: Scaling or standardization involves transforming the data such that it has a mean of 0 and a standard deviation of 1. This process helps prevent variables with larger scales from dominating those with smaller scales during model training. The Python function `StandardScaler()` is leveraged for this task.

e) Splitting the data: The dataset should be divided into training and testing sets using the `train_test_split()` Python function. This partitioning enables

the evaluation of model performance on unseen data, crucial for assessing its generalization capabilities and detecting potential overfitting.

f) Addressing possible class imbalance: To address the imbalance in the target class distribution, the SMOTE oversampling strategy is employed. Synthetic Minority Oversampling Technique (SMOTE) is a sophisticated oversampling strategy that enhances the categorization of minority classes in unbalanced data. SMOTE generates synthetic instances of the minority class by interpolating between existing minority class samples. By oversampling the minority class and balancing the dataset, SMOTE helps improve the performance and reliability of machine learning models (Chawla et al., 2002). For the purposes of this study, we consider a dataset imbalanced if at least one class represents less than 10% of the total dataset, or if the majority class is at least 3 times larger than the smallest class. This is therefore the threshold used to determine the need for applying SMOTE.

3.5.2 Modeling

This paper aims to predict the risk category in which loan-taking customers lie based on various factors including the type of loan acquired and the number of days the borrower has exceeded the agreed repayment date of the funds. There are two experiments conducted in this paper: the first to apply different machine learning algorithms to evaluate which of the algorithms performs the best for predicting the `Risk_Class_Held` variable, and the second where the SMOTE technique was leveraged to mitigate the class imbalance in the variable `Risk_Class_Held`.

A. A Comparison of the Machine Learning Algorithms Predictive Performances

The following machine learning algorithms were employed for the classification task in this paper:

Decision Trees

A decision tree is a supervised learning algorithm, meaning it learns from labeled data where the correct output is already provided. Decision trees are hierarchical, tree-like structures which consist of root nodes, internal nodes, leaf nodes and branches. They are designed to model decisions by recursively splitting the data into subsets based on the most significant feature at each node of the tree. Each node in the tree represents a decision point, where the algorithm selects the feature that best separates the data (Kingsford and Salzberg, 2008). Figure 3.2 illustrates the basic structure of a decision tree.

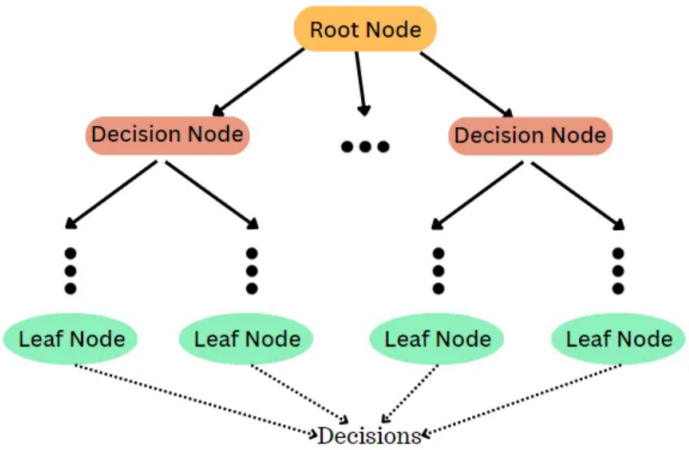


Figure 3.2: Decision Trees
Source: (Shanmugasundaram, 2023)

The algorithm chooses the best feature to divide the data at every node in the tree, with the goal of reducing impurity or maximizing information gain as it creates splits, creating the purest subsets of data. Several attribute selection measures, also known as splitting rules, have been designed to evaluate the degree of impurity, which is the degree of disorder or mixed classes within a set of instances. The most common attribute selection measures are entropy and the Gini index.

Entropy is a measure of disorder; it reflects how mixed the classes are within a data set. It represents the average amount of information needed to determine the class of a data point in the set. An entropy of 0 indicates a perfectly pure set, meaning all data points belong to the same class, and the maximum entropy occurs when all classes are equally represented (all p_i values are equal), representing the highest level of impurity. The formula for entropy is as follows:

$$E = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{3.1}$$

where:

a. E is the entropy of the set E

b. p_i is the fraction of data points in set E that belong to class i

c. The summation is over all classes ($i = 1$ to m)

The Gini Index is a measure that calculates the probability of misclassifying a randomly chosen data point from the set. It reflects how often a randomly selected data point would be incorrectly classified if it were randomly labeled based on the class distribution in the set. Similar to entropy, a Gini index of 0

indicates a perfectly pure set while it approaches 1 when the classes are equally distributed. It's formula is as follows:

$$Gini(D) = 1 - \sum_{j=1}^{c} (p_j)^2 \tag{3.2}$$

where:

a. Gini(D) is the Gini impurity for data set D

b. c is the number of classes

c. p_j is the probability of a sample belonging to class j

The algorithm calculates the entropy or Gini Index for each possible split, evaluating the impurity of the resulting child nodes. It then determines the information gain (when using entropy) or Gini gain (when using Gini Index), which is the the difference in impurity between the parent node and the weighted average impurity of the child nodes. Finally, the algorithm selects the split that yields the highest information gain or Gini gain, as this split leads to the greatest reduction in impurity (Kingsford and Salzberg, 2008). The process of splitting continues until certain conditions are met, such as reaching a specified tree depth or having a minimum number of samples in the leaf nodes. When no further splits are made, a node becomes a leaf, and it is assigned the majority class label.

If decision trees grown unconstrained, they can overfit the training data, capturing noise instead of general patterns. To prevent this, Scikit-Learn provides constraints that help to generalize the model and improve its performance on unseen data (Scikit-Learn, 2024). These include:

a. Maximum Depth (max_depth): which limits how deep the tree can grow.

b. Minimum Samples per Split (min_samples_split): which prevents a node from splitting if it contains fewer samples than this threshold, ensuring meaningful divisions.

c. Minimum Samples per Leaf (min_samples_leaf): which ensures that a leaf node has at least a minimum number of samples, reducing variance.

d. Maximum Number of Features (max_features): which limits the number of features considered for splitting introducing randomness to improve generalization.

Pruning is the process of trimming a decision tree by removing sections that provide little to no gain in predictive power. Pruning techniques further refine decision trees by removing branches that do not contribute significantly to predictive

accuracy. Cost Complexity Pruning (ccp_alpha) is a pruning method which penalizes complex trees by introducing a cost function that balances accuracy and tree size. A higher ccp_alpha value results in a smaller, simpler tree while a lower ccp_alpha value retains more splits, leading to potential overfitting. To further optimize decision tree performance, hyperparameters tuning can be applied using techniques like Grid Search (GridSearchCV) which tests multiple combinations of parameters and Random Search (RandomizedSearchCV) which randomly samples hyperparameters for faster tuning (Scikit-Learn, 2024). These methods help identify the optimal model configuration for better predictive performance.

Random Forests

Random Forests (RF) algorithm represents a method in machine learning that combines the results of several decision trees to make better predictions. This approach, known as ensemble learning (where multiple models are combined to improve performance), enhances prediction accuracy, reliability, and generalization-the ability to work well on new, unseen data. Unlike a single decision tree, which recursively splits the dataset into subsets based on the most important or influential features, the RF algorithm builds many decision trees (forming a 'forest') using random subsets of the data (called bootstrap samples, where data points are randomly selected with replacement) and a random selection of features at each decision point or split. (Breiman, 2001). The purpose of introducing randomness in Random Forest is to reduce the chances of the model being too closely fitted to the training data (a problem known as overfitting), by lowering variation in the predictions made by individual trees. Figure 3.3 illustrates the structure of a Random Forest model.

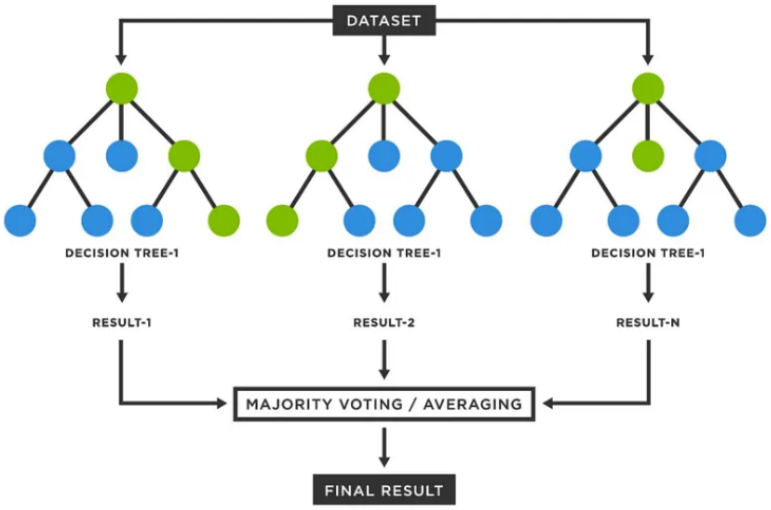


Figure 3.3: Random Forests
Source: (Alam, 2023)

At each node of the tree, the algorithm selects the optimal feature to split the data, aiming to minimize impurity or maximize information gain, typically using measures like Gini impurity or entropy, a measure of randomness or unpredictability in the data set, which is calculated as follows:

$$H(X) = -\sum_{i} P(x_i) \log_2(P(x_i)) \tag{3.3}$$

where:

a. $H(X)$ represents the Shannon entropy of the random variable X

b. $P(x_i)$ denotes the probability of the ith outcome of the random variable X

c. log_2 is the base-2 logarithm.

The splitting process continues until a stopping criterion is met, such as reaching a predefined tree depth or a minimum number of samples in the leaf nodes. According to the Scikit-Learn documentation, the main parameters that influence this process are n_estimators and max_features. The n_estimators parameter controls the number of trees in the forest - generally, a higher number of trees improves performance but increases computational cost, with diminishing returns beyond a certain point. Meanwhile, max_features determines the number of features considered for splitting at each node. Lower values help reduce variance but may introduce higher bias, while higher values allow for more complex splits, potentially increasing variance (Scikit-Learn, 2023). The predictions from individual trees are then aggregated, often through a majority vote for classification tasks or averaging for regression analysis. To optimize the performance of

a Random Forest model, hyperparameter tuning techniques such as Grid Search, Random Search, and the Out-of-Bag (OOB) score can be utilized. The OOB score (oob_score) provides an estimate of model accuracy by evaluating samples that were not included in the bootstrap training process.

Gradient Boosted Decision Trees

Gradient Boosted Decision Trees (GBDT) is a machine learning algorithm that combines the power of decision trees with boosting to improve predictive accuracy. It is an ensemble learning technique that works by successively generating a collection of decision trees where each tree corrects the errors of the previous one, as opposed to Random Forests that create multiple trees independently. The algorithm begins with a single decision tree, which serves as the initial weak learner, and then subsequent trees are fitted, with a focus to the residual errors (the differences between the predicted values and the actual outcomes) made by the previous trees. The process is iterated until a predetermined number of trees or a specified level of performance is attained. At each iteration, the algorithm assigns weights to each tree based on their contribution to minimizing the overall loss function, which is a measure of how far the model's predictions are from the true values. This is typically done using techniques like gradient descent, a method for gradually adjusting the model to reduce the loss function step by step. The final prediction is then made by combining the predictions of all the trees in the ensemble. This iterative process allows GBDT to continuously improve its predictions by focusing on the mistakes made by the previous trees (Friedman, 2001).

For this paper, two implementations of GBDT were leveraged to perform modeling:

XGBoost (Extreme Gradient Boosting)

XGBoost is a powerful ensemble model known for its exceptional performance in various machine learning applications. It is an advanced form of gradient boosting, which sequentially combines the results of weaker models (typically decision trees) to form a strong predictive model. XGBoost minimizes a loss function by adding decision trees iteratively. Each new tree corrects the errors made by the previous ones (Chen et al., 2015). The algorithm leverages a technique known as regularized learning, which employs L1 (Lasso) and L2 (Ridge) regularization techniques, and is used to prevent overfitting and enhance the model's ability to generalize to unseen data (Chen et al., 2015). XGBoost

aims to minimize the the following objective function:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{3.4}$$

The equation consists of the following main components:

a. Loss Function - The difference between the predicted output ($\hat{y}_i$) and the target value (y_i).

b. Regularization Term - Added to the loss function to penalize (control) the model's complexity that is represented by the tree functions (f_k), to prevent overfitting.

c. Regularization Function - Defines how the complexity penalty is calculated, and is defined as $\Omega(f)$

Where:

a. $\phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$ represents the model's prediction for instance i using K additive functions (trees).

b. $\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2$ defines the regularization term, with γ controlling the penalty on the number of leaves (T) and λ controlling the penalty on the magnitude of leaf weights (w).

XGBoost similarly leverages Gradient Tree Boosting which involves adding a new tree (f_t) to the model to minimize the loss where $\hat{y}_i^{(t-1)}$ is the prediction for instance i at the previous iteration (Chen et al., 2015). Here, the (f_t) that most improves the model is greedily added.

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3.5}$$

To efficiently optimize the objective, XGBoost utilizes a second-order Taylor approximation as follows:

$$L^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{3.6}$$

where $g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$ represents the first-order gradient statistic and $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}$ represents the second-order gradient statistic.

The constant terms are removed, and rewritten by expanding Ω, obtaining the following simplified objective at step t.

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{3.7}$$

where I_j represents the set of instances belonging to leaf j. This equation defines a score to evaluate the quality of a tree structure (q), similar to the impurity score for evaluating decision trees.

XGBoost expands all nodes at the same depth before moving deeper. The loss reduction function, or split gain, after splitting a node into two child nodes is given by

$$L_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{3.8}$$

where I_L and I_R represent the sets of instances belonging to the left and right child nodes, respectively. This equation is utilised for evaluating the split candidates.

To optimize the algorithms performance, several hyperparameters can be tuned including:

a. n_estimators - This specifies the number of boosting rounds or trees to be added. While increasing this can improve model performance, it may also lead to overfitting.

b. max_depth - This determines the maximum depth of each tree. Deeper trees can capture more information but might overfit; shallower trees may underfit.

c. learning_rate (eta) - This controls the contribution of each tree. Lower values require more trees but can lead to better generalization.

LightGBM (Light Gradient Boosting Machine)

LightGBM is a high-performance gradient boosting framework that excels in training large datasets with high efficiency. Similar to XGBoost, it employs an ensemble of decision trees to progressively refine predictions, optimizing a chosen loss function through iterative tree-based learning. LightGBM stands out for its speed and scalability, making it particularly well-suited for handling massive datasets (Ke et al., 2017). LightGBM follows a similar approach to XGBoost in minimizing a general objective function. However, unlike XGBoost's level-wise

tree growth, LightGBM employs a leaf-wise (best-first) growth strategy, selecting the leaf with the maximum delta loss to expand. Figure 3.4 visually demonstrates how LightGBM grows trees leaf-wise by expanding the most promising leaf at each step. This approach improves efficiency, increases training speed, and achieves lower loss, especially for large datasets.

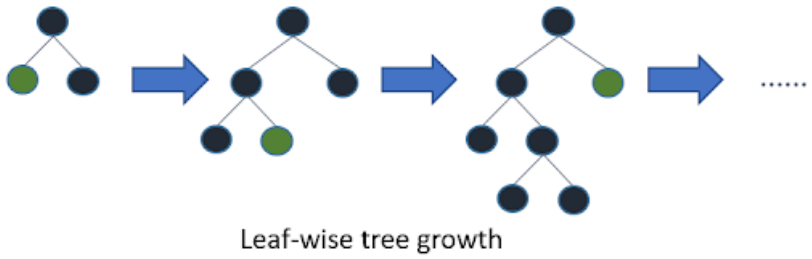


Figure 3.4: Leaf-Wise Tree Growth
Source: (LightGBM, 2025)

The goal of the algorithm is to minimize a loss function $L(y, \hat{y})$ iteratively. The objective function consists of two parts:

$$O = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \Omega(T) \tag{3.9}$$

where:

a. $L(y_i, \hat{y}_i)$ is the loss function (e.g., Log Loss for classification, MSE for regression).

b. $\Omega(T)$ is the regularization term to control model complexity.

Each boosting iteration computes the first-order gradient (g) and second-order Hessian (h) to approximate the loss function:

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}, \quad h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \tag{3.10}$$

For log loss, they are computed as:

$$g_i = \sigma(\hat{y}_i) - y_i, \quad h_i = \sigma(\hat{y}_i)(1 - \sigma(\hat{y}_i)) \tag{3.11}$$

These values are used to compute split gain and update leaf values. The gain for splitting a node is calculated as:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{3.12}$$

where:

a. G_L, G_R are the sum of gradients for the left and right child nodes.

b. H_L, H_R are the sum of Hessians for the left and right child nodes.

c. λ is the regularization parameter (L2 penalty).

d. γ is the complexity penalty for adding a new leaf.

A split is chosen if Gain > 0.

Light GBM leverages two novel techniques to decide the best feature split:

(a) Gradient-based One-Side Sampling (GOSS): This technique addresses the absence of native instance weights in GBDT. It recognizes that data instances have varying weights and differing impacts on information gain computation. By definition, instances with larger gradients contribute more to information gain. Therefore, when downsampling instances, it's optimal to retain those with significant gradients and randomly discard only those with small gradients. GOSS keeps all instances with large gradients while randomly sampling instances with smaller gradients. This approach improves the accuracy of gain estimation compared to uniform random sampling (Ke et al., 2017).

(b) Exclusive Feature Bundling (EFB): This is a technique that is used to simplify datasets by reducing the number of features in the dataset without losing much information. This is done by grouping together features that tend to appear separately. EFB attains this by bundling exclusive features thus retaining fewer groups that represent similar information. The technique also ensures that features within each group do not tend to appear together frequently. This optimization process helps streamline the dataset while maintaining important distinctions between features (Ke et al., 2017).

Since LightGBM uses a leaf-wise growth approach, it requires proper regularization to prevent overfitting. The main parameters to control model complexity include:

a. Max Depth (max_depth): Limits the depth of trees to avoid excessive growth.

b. Number of Leaves (num_leaves): Controls the complexity of the trees. More leaves allow finer splits but may lead to overfitting.

c. **L1 & L2 Regularization (lambda_l1, lambda_l2)**: Helps reduce overfitting by penalizing large coefficients.

d. **Min Data in Leaf (min_data_in_leaf)**: Ensures each leaf has a minimum number of data points, preventing small, highly specific leaves.

e. **Number of Trees (n_estimators)**: Defines the number of trees in the ensemble. More trees generally improve performance but increase computation time.

f. **Learning Rate (learning_rate)**: Determines how much the model adjusts after each iteration. Lower values improve generalization but require more trees to converge.

To achieve the best performance, hyperparameter tuning techniques like Grid Search, Random Search, and Bayesian Optimization can be used. Additionally, LightGBM provides an Out-of-Bag (OOB) Score (bagging_fraction and bagging_freq), which evaluates model accuracy using samples not included in the bootstrap training process (LightGBM, 2025).

B. Evaluating the Impact of Using the Oversampling Technique SMOTE on ML Algorithms to Address Class Imbalance

Many real-world classification problems exhibit imbalanced class distributions, presenting a significant challenge for machine learning algorithms. The problem that lies in class imbalance learning is that standard methods often struggle with misclassifying positive class samples as negative class samples. Additionally, these methods face difficulty learning the minority class due to its limited representation in the dataset. To overcome this challenge, various techniques have been developed, such as oversampling. Oversampling is a technique which involves randomly duplicating minority class samples to increase their prevalence in the dataset. However, traditional oversampling methods may lead to overfitting. To address this problem, SMOTE was introduced to enhance random oversampling by generating synthetic minority class samples thus balancing the class distribution (Raghuwanshi and Shukla, 2020). Hence, to mitigate potential class distribution imbalances in the dataset utilized for this project, the informed oversampling approach, SMOTE, is employed.

Synthetic Minority Oversampling Technique (SMOTE) is a sophisticated oversampling strategy that enhances the categorization of minority classes in unbalanced data. SMOTE generates synthetic instances of the minority class rather

than performing over-sampling with replacement. The SMOTE technique involves oversampling the minority class by taking each minority class sample and generating synthetic examples based on the k-nearest neighbors of each minority class sample. The synthetic samples themselves are generated by interpolating between existing minority class samples as follows: the difference between the feature vectors (sample) under consideration and its nearest neighbor is multiplied by a random number between 0 and 1, and then added to the feature vector. This process leads to the selection of a random point along the line segment joining two specific features, resulting in the creation of synthetic examples that encourage the classifier to construct larger and less specific decision regions. Consequently, classifiers are better able to generalize, as they learn more general regions for minority class samples rather than being influenced solely by the majority class samples surrounding them (Chawla et al., 2002). Therefore, by oversampling the minority classes and balancing the dataset, SMOTE aids in improving the performance and reliability of the machine learning model utilized in this paper.

The SMOTE over-sampling algorithm is outlined as follows:

Step 1: Set the minority class set A. For each random $x \in A$, identify the k-nearest neighbors of x, obtained by calculating the Euclidean distance between x and every other sample in set A.

Step 2: Randomly select N examples from the k-nearest neighbors of x, constructing the set A_1. Select a random sample x_k from the set, where k is the rank of the neighbor (i.e $k = 1, 2, 3, \ldots, N$).

Step 3: Perform linear interpolation between x and x_k to generate a new example as follows:

$$x' = x + w \cdot |x - x_k| \tag{3.13}$$

where w = rand$(0, 1)$, which represents a random number between 0 and 1.

Step 4: Repeat step 1 to 3 until M synthetic samples are generated.

Figure 3.5 provides a visual representation of SMOTE interpolation displaying original minority samples and a SMOTE generated sample.

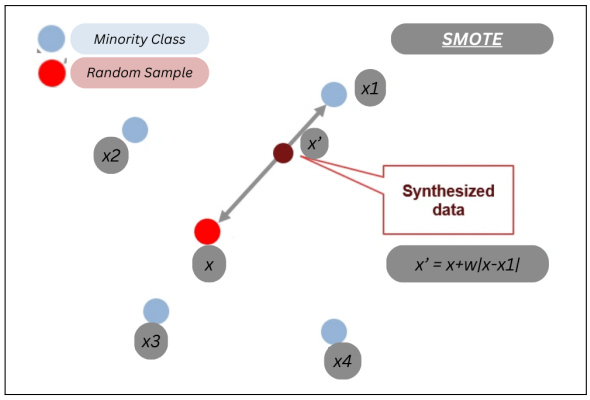


Figure 3.5: SMOTE Over-Sampling Algorithm

Although standard SMOTE effectively mitigates class imbalance, researchers have developed several improved versions to address its limitations, particularly overgeneralization, oversampling noise, and class boundary issues. For instance, Borderline-SMOTE focuses on oversampling minority class instances near the decision boundary, enhancing class separability, while ADASYN (Adaptive Synthetic Sampling) assigns higher oversampling rates to harder-to-classify samples, improving the model's ability to distinguish between classes (Han et al., 2005; He et al., 2008). Despite these refinements, this study employs standard SMOTE due to its simplicity, effectiveness, and widespread adoption in handling imbalanced datasets. The primary focus is to evaluate the impact of oversampling on classification performance before considering more complex variations.

3.5.3 Evaluation

Cross-Validation

Cross-Validation is a technique that is used in machine learning for model performance evaluation. The technique is used to check how well a machine learning model will perform on new, unseen data. It is widely used to compare different machine learning models and determine the optimal one. The technique offers two key benefits in these experiments:

1. Prevention of overfitting; A situation where the model performs well on the training data but poorly on new, unseen data because it has become too specialized to the training set.

2. Reduction of noise sensitivity - It minimizes the model's tendency to learn noise in the data, which can be worsened by techniques like SMOTE.

These benefits ensure that the model generalizes well to new, unseen data and avoids the learning of noise.

There are various types of cross-validation techniques, with k-Fold Cross-Validation, Leave-One-Out Cross-Validation (LOOCV) and Hold-Out Cross-Validation being among the most common types. This paper utilizes Stratified k-Fold Cross-Validation, a variation of k-Fold Cross-Validation that is particularly useful in handling class imbalance in datasets. The technique works as follows:

1. Dividing the Dataset: The dataset is divided into k equal-sized subsets. Stratified k-Fold ensures that each fold maintains the same class distribution as the entire dataset (unlike standard k-Fold where splitting is random and class balance can be uneven across folds), which prevents biased training and testing, a crucial step for handling imbalanced datasets.
2. Training and Testing the Model: The model uses $k - 1$ folds as the training set and the remaining fold as the test set.
3. Calculating Performance Metrics: Metrics, e.g., accuracy, precision, recall, F1-score, are calculated and recorded.
4. Iterating: The process is repeated k times.
5. Averaging of Results: The final model evaluation is done by calculating the average of all k iterations.

Figure 3.6 below illustrates the process of Stratified k-Fold Cross-Validation.

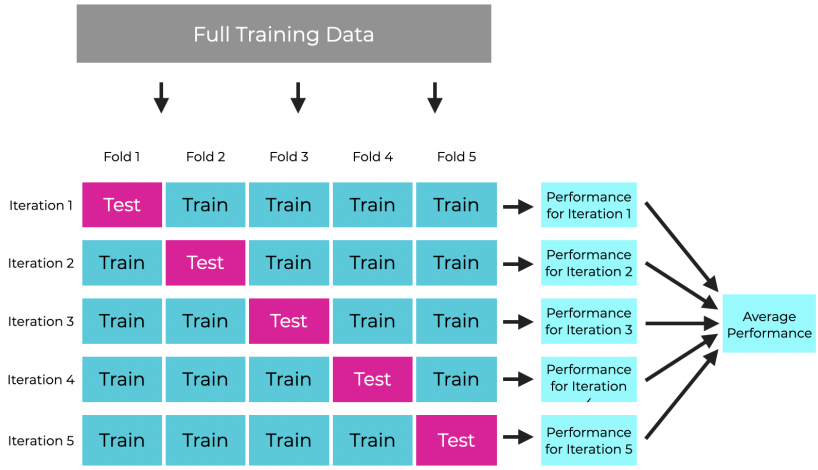


Figure 3.6: k-Fold Cross-Validation
Source: (Lumumba et al., 2024)

Confusion Matrix

The confusion matrix is a simple yet effective tool in evaluating the performance of supervised machine learning models classification tasks. It presents a tabular represen-

tation (square matrix) of predicted classes against actual classes as shown in table 3.2, highlighting the number of correct and incorrect predictions made by the model. This matrix enables quick assessment of model accuracy, precision, recall, and other performance metrics. The metrics are defined as follows:

a. Accuracy: The percentage of values that are correctly categorized. It indicates how often the classifier is correct.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.14}$$

b. Precision: Precision determines the model's ability to correctly classify positive values.

$$Precision = \frac{TP}{TP + FP} \tag{3.15}$$

c. Recall: This score is used to determine the model's predictive performance for true positive values.

$$Recall = \frac{TP}{TP + FN} \tag{3.16}$$

d. F1-score: It is the harmonic mean of precision and recall, providing a single numerical value that balances the trade-off between these two measures.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.17}$$

where:

a. True Negative (TN): The number of (Actual) negatives correctly classified as negatives by the model classification.

b. False Negative (FN): The number of (Actual) positives that are incorrectly labeled as negatives.

c. True Positive (TP): The number of accurately classified (Actual) positives.

d. False Positive (FP): The number of (Actual) negatives that are incorrectly identified as positives by the model.

Insights into a model's accuracy, precision, recall, and general efficacy in classifying instances are provided by the analysis of true positive, true negative, false positive,

and false negative predictions it offers. For a multi-class problem, we consider each class, treating them individually, as there is no single ”positive” or ”negative” class as in binary classification. In table 3.2, the calculation for the values of Class 1 is shown. True Positives (TP) refer to the count of instances correctly identified as belonging to Class 1, aligning with the actual target values. True Negatives (TN) represent the total number of instances from classes other than Class 1 that were correctly predicted as not being Class 1. False Positives (FP) are the instances from other classes that were incorrectly predicted as Class 1. False Negatives (FN) are the instances that should have been identified as Class 1 but were mistakenly predicted as belonging to other classes. This process is repeated for all classes, and the overall performance summarized using macro (average per class) or micro (overall counts) averaging methods.

Table 3.2: Multi-Class Confusion Matrix

Actual	**Predicted**		
	Class 1	**Class 2**	**Class 3**
Class 1	TP	FN	FN
Class 2	FP	TN	TN
Class 3	FP	TN	TN

3.6 Deployment

The deployment of an application tailored to multi-class credit risk classification within lending institutions was executed by leveraging the Visual Studio Code (VS Code) environment, to design the user interface (UI) to ensure intuitive interaction and seamless functionality. With a pre-trained and saved model, carefully selected based on its superior performance, the application attained predictive capabilities. This model was seamlessly integrated into the main application code, for use in real-time credit risk predictions. Subsequently, the application was deployed on the Streamlit platform, offering accessibility and ease of use via a web-based interface.

4. System Design and Architecture

4.1 Introduction

This chapter offers an overview of the critical aspects involved in designing the system and architecture necessary for developing and deploying an effective multi-class credit risk prediction model. With a focus on system design and architecture, the aim of this chapter is to provide insight into the systematic approach employed to construct the application architecture that leverages the prediction model.

4.2 Overview of the System Architecture, Design and Components Interaction

Developing the credit risk prediction application involved a streamlined process that revolved around designing a user-friendly Streamlit application. The core structure of the project consisted of a model for predicting multi-class credit risk and a Streamlit user interface (UI) to interact with the model. The initial step involved retrieving the dataset, which was stored in a flat file format (.xlsx). Python's Pandas library facilitated the retrieval process, allowing seamless access to the dataset within the Python environment. Once the dataset was retrieved, preprocessing was conducted to ensure the quality and integrity of the data for model training. This entailed encoding categorical variables, and scaling numerical features. This was achieved by leveraging Pandas and NumPy libraries, setting the stage for effective model training. With the preprocessed dataset in hand, the next step focused on developing the credit risk prediction model. Modeling was performed on both the imbalanced dataset and the balanced datset which was generated by applying the SMOTE technique onto the data. The model, constructed using the XGBoost algorithm, which was the best performing algorithm of the four tested algorithms, was trained to predict credit risk based on the input features. Once the model was trained and validated, Python scripts, orchestrated within the Visual Studio Code (VS Code) integrated development environment (IDE), together with VS Codes integration with Streamlit were leveraged for the deployment and saving of the interactive application used for multi-class credit risk scoring. The Streamlit application UI, designed to seamlessly integrate with the underlying predictive model, provides users with an intuitive platform to input borrower information and loan details. Upon input submission, the application calls the model to generate real-time predictions on credit risk classes, empowering lenders to make informed de-

cisions regarding loan approvals and risk management strategies. In summary, the implementation process encapsulated data retrieval, preprocessing, model development, evaluation and selection, and UI design and deployment, culminating in the creation of a robust multi-class credit risk prediction application. By adopting a systematic approach and leveraging the capabilities of Python, Streamlit, and VS Code, the project successfully translated the system design into a practical solution that effectively analyses multi-class credit risk. fig. 4.1 illustrates the overview of the system components, architecture and design.

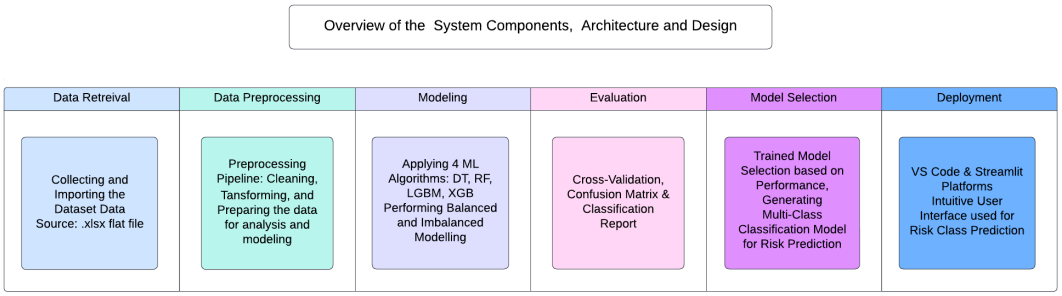


Figure 4.1: Overview of the System Components, Architecture and Design

4.3 System Implementation and Testing

The implementation phase marked a pivotal moment where theoretical concepts materialized into tangible solutions. The implementation phase commenced with the development of the user interface (UI) using the Streamlit framework. Leveraging Streamlit's intuitive API and Python's flexibility, the UI was designed to provide a seamless experience for end-users. The UI was crafted to capture borrower information and loan details through interactive input fields, ensuring ease of use and accessibility. Once the UI was developed, the next step involved integrating it with the underlying predictive model. The trained model, developed using the XGBoost algorithm, was incorporated into the Streamlit application to enable real-time credit risk predictions. Python scripts were utilized to establish communication between the UI and the model, allowing for efficient data exchange and prediction generation. A structured testing protocol was devised to assess various aspects of the system's UI, including navigation and input validation. Maintenance and support of the system will be conducted by regular monitoring of the deployed system to identify and address any issues or performance bottlenecks. Below is a snapshot illustrating the functionality of the application.

The image in fig. 4.2 encapsulates the core functionality of the application, highlighting its ability to empower lenders with timely and accurate risk assessments to inform their decision-making process. It showcases the user-friendly interface of the credit risk

prediction application, illustrating the seamless interaction between the input fields and the predictive model. Users can input borrower information and loan details with ease, facilitated by intuitive input fields.

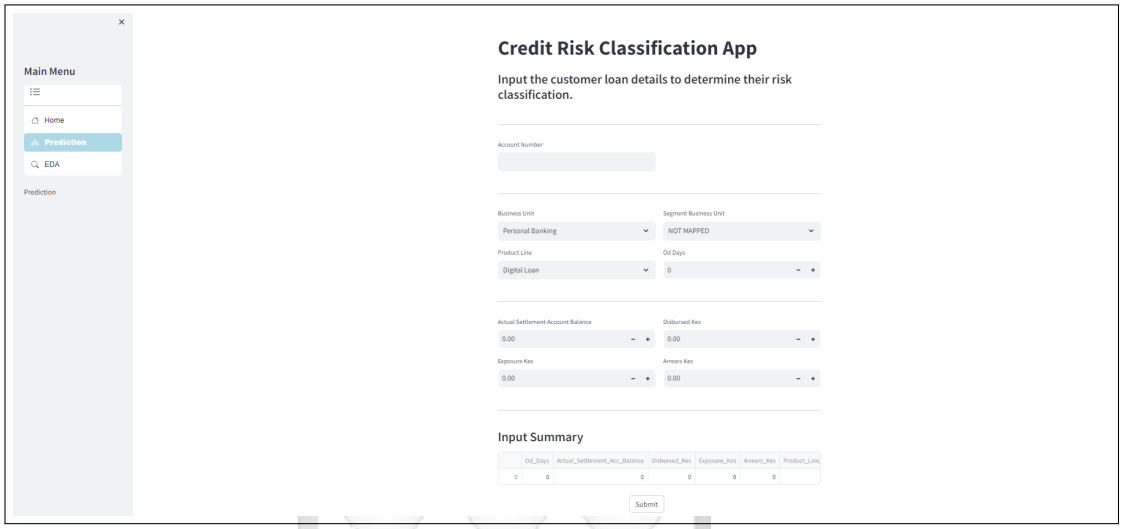


Figure 4.2: Application User Interface

Upon submission, the application swiftly processes the input data and generates real-time predictions on the credit risk class, providing users with valuable insights into the risk profile of the loan as shown in fig. 4.3.

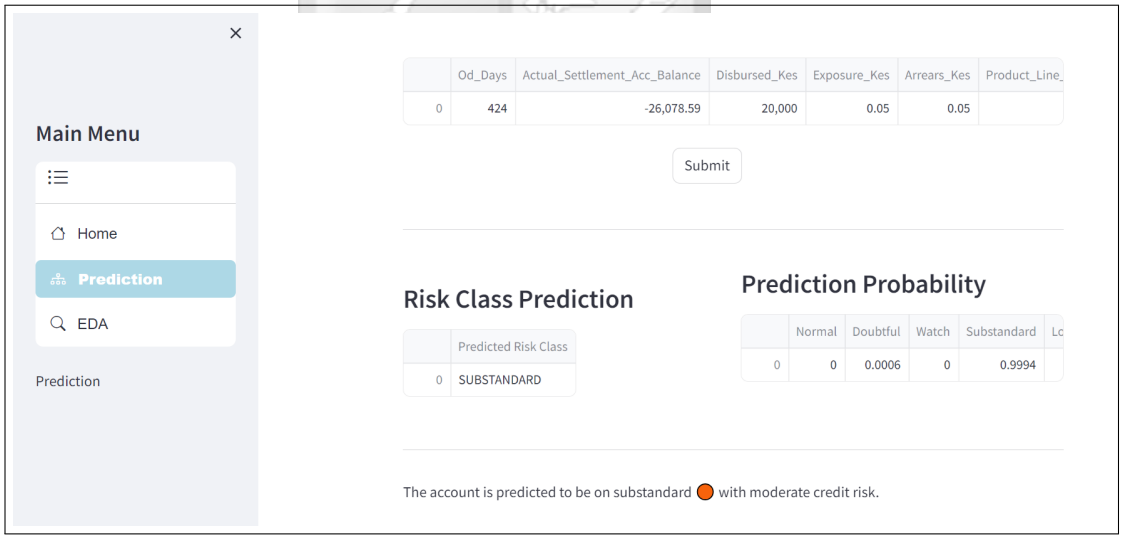


Figure 4.3: Application Risk Class Prediction

4.4 Conclusion

In conclusion, the systematic design and implementation of the credit risk prediction application were successfully executed, culminating in the deployment of a robust

solution for lenders. By leveraging Python, Streamlit, and Visual Studio Code, the project navigated through the complexities of data retrieval, preprocessing, model development, and user interface design. The seamless integration of these components resulted in a user-friendly application for multi-class credit risk assessment and prediction that empowers lenders to assess credit risk in real-time and make informed decisions regarding loan approvals. Moving forward, ongoing maintenance and support efforts will sustain the effectiveness and longevity of the deployed system, contributing to improved risk management practices in the lending industry. The next section delves into the results chapter where discussion of key results are outlined.

5. Discussion of Results

5.1 Introduction

This section presents comprehensive results of the analysis of the dataset and the machine learning modeling, aiming to provide insights and findings derived from the outcomes of the experiments stated and conducted in the methodology section. The main objective of this paper was to develop and implement an effective credit risk prediction model tailored to the lending landscape in Kenya, by utilizing a multi-class classification technique to enhance loan portfolio management in Kenyan lending institutions. This study aimed to propose a model that seeks to contribute to the advancement of loan portfolio and credit risk management practices in the Kenyan financial sector, ultimately leading to more informed decision-making processes and improved financial performance. To achieve this goal, various analytical techniques were employed to explore the dataset, identify patterns, and uncover relationships between different variables. Further, various modeling algorithms have been explored and modeled on the data in a bid to realize the most promising model for this task in order to achieve this objective.

5.2 Data Preparation

5.2.1 Data Cleaning

1. Handling missing values: Missing values in the column `Actual_Settlement_Acc_Balance` totaled 8. These missing values were replaced with '0'. Additionally, the columns `Restructure_Reason` and `Aa_Problem_Category` had a large proportion of missing values, 91.38% and 99.82% respectively. Due to their high percentage of missing data, these two columns were dropped from the dataset as they did not appear to provide useful information for analysis or modeling, potentially impacting the accuracy of results. As a result of these actions, the dataset contained no remaining null values.

2. Treating duplicate data: The dataset contained only 2 duplicates which were dropped leaving the dataset with no duplicates.

3. Correcting data types: The columns `Account_Officer_Code` and `Branch_Code` were converted to strings data type. This adjustment ensured that these columns

were treated as categorical variables rather than numerical ones, which was beneficial for subsequent analysis.

4. Dropping Columns: Several columns, including `Currency`, `Exposure_Actual`, `Arrears_Actual`, `Account_Officer_Code`, and `Branch_Code`, were removed from the dataset due to their redundancy or lack of relevance for the analysis. After dropping these columns, the dataset was left with a total of 10 remaining columns, ensuring a more streamlined and focused dataset for subsequent modeling and analysis tasks.

5.2.2 Exploratory Data Analysis

Univariate Data Analysis

The target variable, "`Risk Class Held`", comprises five distinct risk classes that indicate the likelihood of a loan defaulting. These categories are defined as follows:

a. Normal: Loans that are up to date with repayments and have minimal risk of default.

b. Watch: Loans that show early signs of potential repayment issues but are not yet considered high-risk.

c. Substandard: Loans with missed payments or financial difficulties, indicating a moderate risk of default.

d. Doubtful: Loans with significant repayment delays and a high likelihood of default, though some recovery may still be possible.

e. Loss: Loans that are unlikely to be repaid and are considered as defaults, with minimal chances of recovery.

Among these classes, 'normal' appears to be the most prevalent, with a count of 9662 instances, followed by 'doubtful', 'watch', 'substandard', and 'loss', with counts of 3744, 2448, 2094, and 1947, respectively. This distribution, as seen in fig. 5.1, highlights the varying degrees of credit risk present in the dataset, with a majority of instances falling within the 'normal' category, indicating relatively low risk.

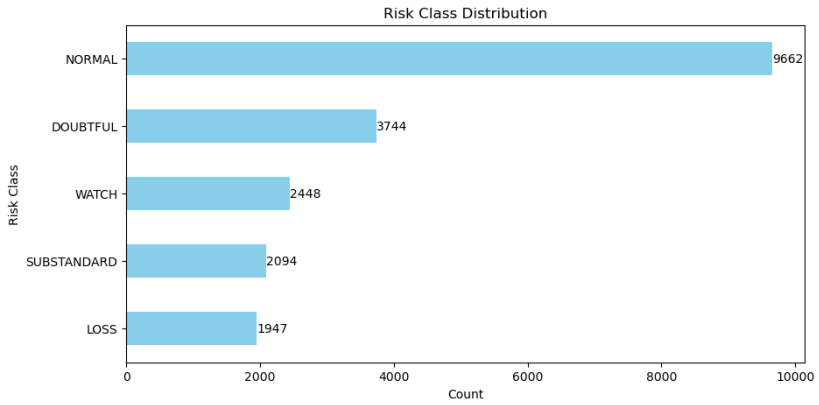


Figure 5.1: Univariate Analysis-Risk Class Distribution

As seen in fig. 5.2, the highest number of loans disbursed belongs to the Digital Loans category, reflecting the significant adoption of digital products among the lender's client base and a reflection of borrowing tendencies within the country as a whole. The second highest category is Asset Finance, suggesting that a substantial percentage of borrowers are seeking loans to acquire physical assets. Figure 5.2 further indicates that holders of personal and business bank accounts are the primary borrowers from the lender, likely due to the convenience and tailored financial products available to them, which address their specific borrowing needs, such as instant digital loans, the most commonly acquired type of loan.

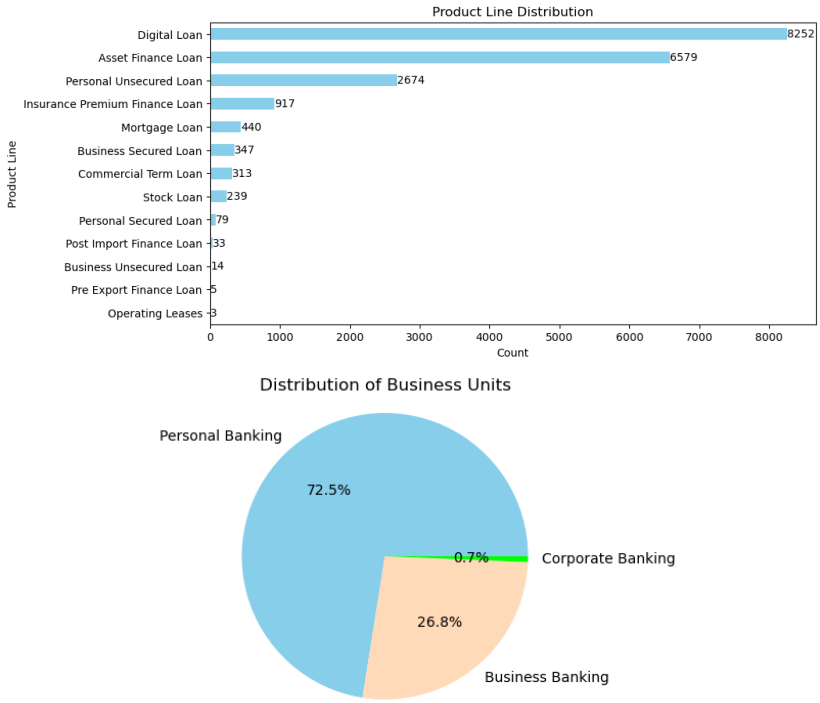


Figure 5.2: Univariate Analysis of Product Line and Business Units Distribution

Bivariate Data Analysis

The bivariate analysis of risk class held by certain categorical features reveals some useful insights. Figure 5.3 illustrates key aspects of these relationships. Regarding the `Business Unit` feature, 'Personal Banking' category exhibited the highest number of normal loans. In the `Segment of Business Unit` column, 'Go Banking' category recorded the highest number of normal class loans along with 'SMEB' and 'Business Banking', while the 'Not Mapped' category showed the highest count of doubtful and loss loans. In terms of the `Product Line` feature, the 'Asset Finance Loan' category had the highest number of normal class loans, whereas the 'Digital Loans' category showed the highest counts for both doubtful and loss classes.

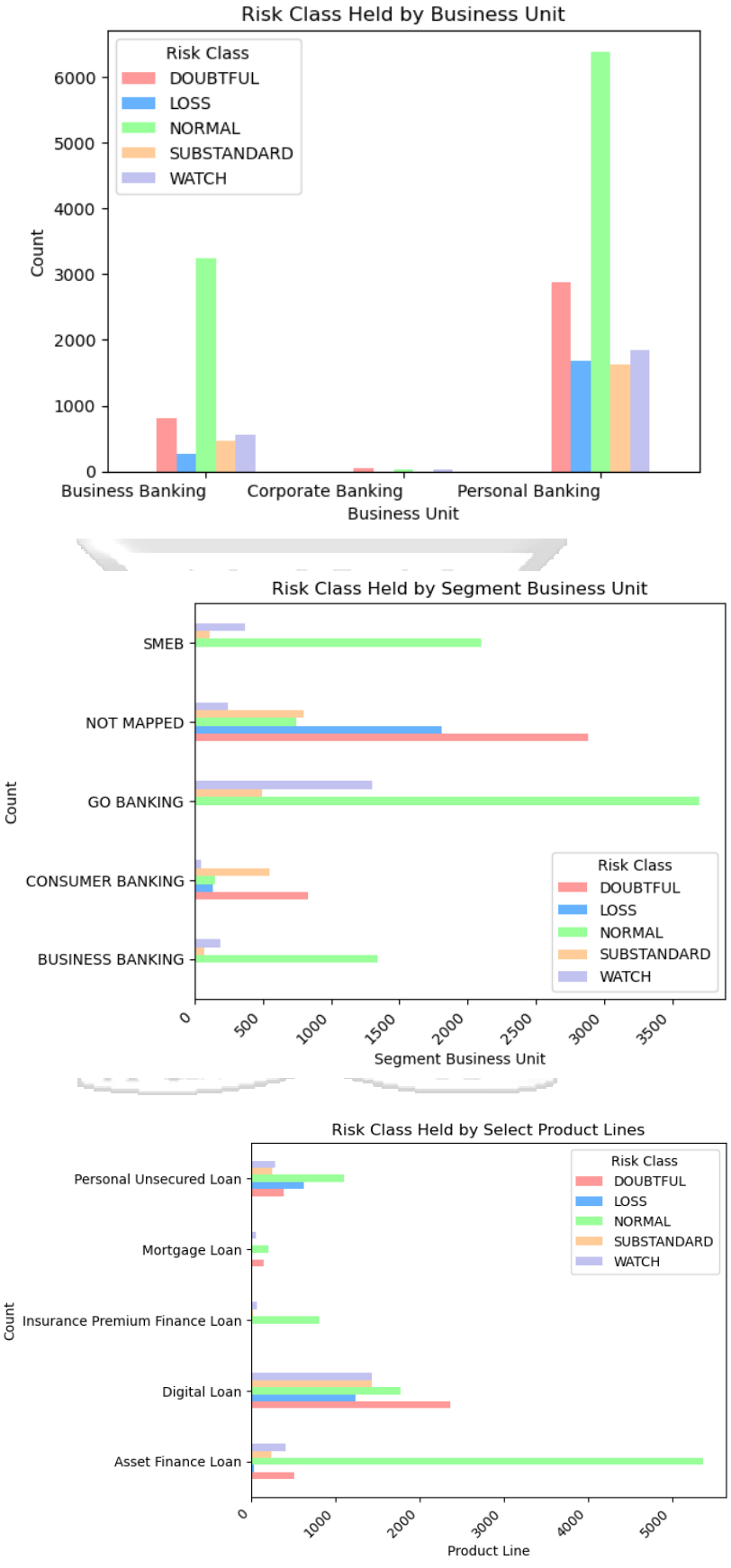


Figure 5.3: Bivariate Analysis - Risk Class by Categorical Columns

Within the `Business Unit` feature, the prevalence of normal loans in the 'Personal Banking' category suggests a lower risk profile associated with these customers. The analysis of the `Segment of Business Unit` column reveals that 'Go Banking', 'SMEB', and 'Business Banking' predominantly consist of normal loans, indicating strong creditworthiness. 'Go Banking' offers current accounts with no ledger fees thus appealing to individual customers while 'SMEB', and 'Business Banking' focus on businesses. The prevalence of normal loans in these segments suggests that both individual customers and business owners are managing their loan repayments well, possibly by securing loans within their repayment capacities. The presence of 'doubtful' and 'loss' loans in 'Not Mapped' raises concerns about data quality and risk assessment, as these loans lack proper classification. Within the `Product Line` feature, the prevalence of normal loans in 'Asset Finance Loan' category suggests a lower risk level, possibly attributed to more stringent eligibility criteria or collateral requirements associated with asset-backed financing. Conversely, the higher incidence of doubtful and loss instances in the 'Digital Loans' category indicates elevated credit risk, likely stemming from factors such as unsecured lending, higher default rates, or challenges in assessing borrower creditworthiness.

As depicted in fig. 5.4, the loss and doubtful classes exhibit the highest number of days for both maximum and average overdue periods, aligning with expectations.

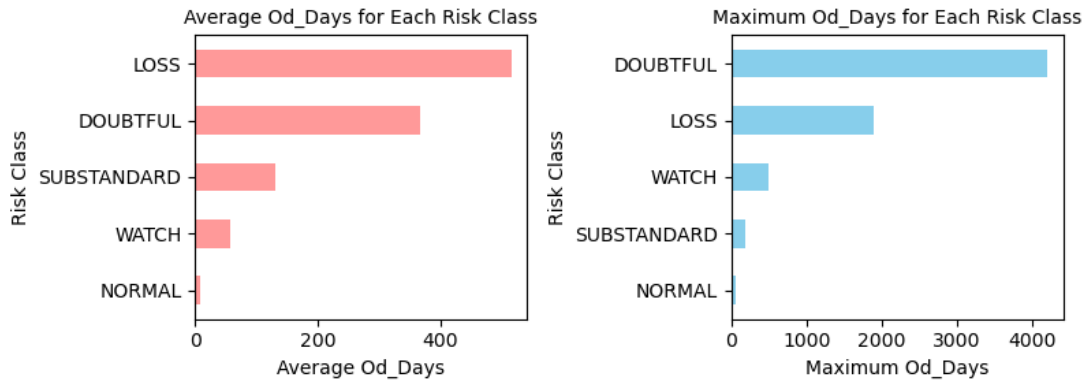


Figure 5.4: Bivariate Analysis - Average & Maximum OD Days

Multivariate Data Analysis

Figure 5.5 displays a correlation heatmap showing the the strength and direction of linear relationships between all pairs of variables.

a. `Actual_Settlement_Acc_Balance` has a negative correlation with `Arrears_Kes` (-0.18) and `Od_Days` (-0.19), suggesting that as the actual settlement account balance decreases, the arrears amount and number of days overdue increase.

b. `Arrears_Kes` has a moderate positive correlation with `Exposure_Kes` (0.61), indicating that higher exposure amounts tend to be associated with higher arrears amounts.

c. `Od_Days` shows a weak positive correlation with the `Disbursed_Kes` and `Exposure_Kes` (0.04), indicating a slight tendency for higher disbursed and exposure amounts to be associated with more overdue days, although the correlation is relatively weak.

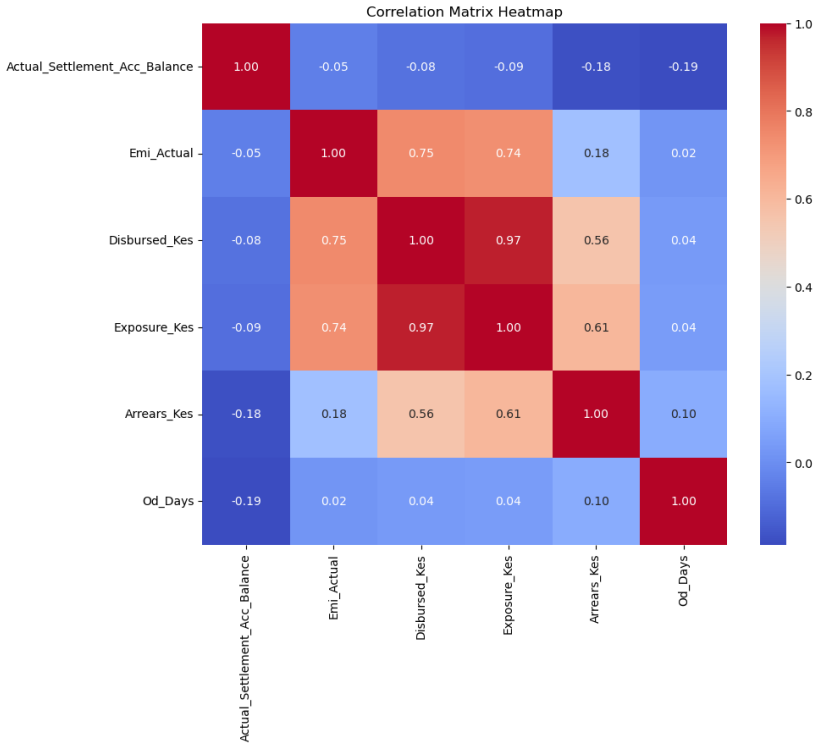


Figure 5.5: Multivariate Analysis

5.3 Machine Learning Modeling

5.3.1 Data Preprocessing

The columns `Account_Officer_Code` and `Branch_Code` were dropped due to their high cardinality, with 640 and 100 unique values respectively. One-hot encoding was applied to the remaining categorical variables to ensure compatibility with the machine learning models. The newly created variables were then concatenated to the original dataset and the original categorical columns dropped. The dataset was then partitioned into predictor variables ($\mathbf{X}$) and the target variable ($\mathbf{y}$). Features ($\mathbf{X}$) comprise all columns except the target variable `Risk_Class_Held`, while the target variable ($\mathbf{y}$) is isolated for classification purposes. Scaling of the numerical variables

was then performed. This was performed by standardizing numerical variables in order to ensure uniformity in their scales. The dataset was then split with a test size of 0.25, allocating 75% of the data for training machine learning models. The split was stratified based on the target variable `Risk_Class_Held` to maintain a proportional class distribution in both sets, preserving the original class distribution.

5.3.2 Modeling and Evaluation

The results of the modeling section are divided into two subsections in line with the two experiments conducted: to evaluate which of the algorithms performs the best for predicting the `Risk_Class_Held` variable and to mitigate the class imbalance in the variable `Risk_Class_Held` by leveraging the SMOTE technique.

A. A Comparison of the Machine Learning Algorithms Predictive Performances

The first experiment described in this paper involved applying various machine learning algorithms to the data in order to compare and evaluate their predictive power. This investigation aimed to determine which algorithm performs best for predicting the `Risk_Class_Held` variable. Here, four algorithms were used i.e. Decision Trees (DT), Random Forests (RF), Extreme Gradient Boosting (XGB) and Light Gradient Boosting Machine (LightGBM). The performance of the four ML algorithms was compared using a stratified 10-Fold Cross-Validation to determine the optimal predictive model. Table 5.1 gives us an overall summary of the prediction performance of the different classifiers on the dataset.

Table 5.1: Overall Metrics for each Model

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.9306	0.9089	0.9105	0.9096
Random Forest	0.9508	0.9469	0.9228	0.9340
LightGBM	0.9503	0.9445	0.9239	0.9337
XGBoost	0.9534	0.9479	0.9285	0.9377

From the results above, we find that all models, except the Decision Tree have similar accuracies around 95%, indicating good overall performance, robustness and effectiveness in classifying the data. The ensemble methods (RF, LightGBM, and XGB), which combine multiple decision trees, provided more consistent performance, likely due to their ability to reduce overfitting compared to the Decision Tree. Based on the results, the Extreme Gradient Boosting model (XGBoost) is the overall best performing model. It achieved the highest accuracy among the models listed, and while Random Forest and LightGBM have al-

most similar performance, XGBoost edges them out slightly once more in terms of precision, producing fewer false positives compared to the others, and recall, capturing a higher proportion of actual positive cases. Overall, XGBoost is the best-performing model, consistently outperforming the other models across all metrics. It achieves the highest accuracy (0.9534), precision (0.9479), recall (0.9285), and F1-score (0.9377). Random Forest and LightGBM are the next best performing algorithms, and are very close in performance. Overall, the three ensemble methods demonstrate their robustness in handling the dataset and outperform traditional Decision Trees for this task. Further analysis on the top-performing model demonstrates how well it distinguishes between the five risk categories. The confusion matrix for XGBoost is presented in fig. 5.6.

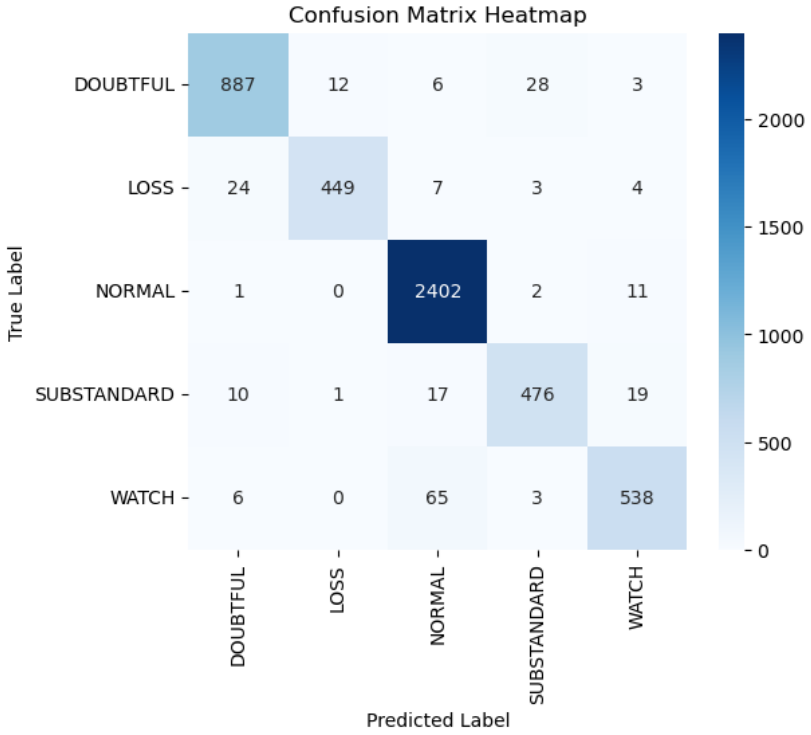


Figure 5.6: XGBoost Confusion Matrix

XGBoost demonstrates strong classification performance for the majority class, Normal, correctly identifying 2,402 instances with minimal errors. It also performs well on Doubtful, the most represented minority class, accurately classifying 887 instances. However, the model exhibits slightly higher misclassification rates for the other three minority classes, with the Watch category experiencing the highest misclassification rate - 65 instances incorrectly classified as Normal. Additionally, misclassification trends reveal challenges in distinguishing borderline cases, particularly where risk categories share overlapping characteristics. Notably, Substandard and Watch loans, Doubtful and Substandard loans, as well

as Loss and Doubtful loans, exhibit classification overlap, suggesting that these categories may have similar risk profiles. While XGBoost performs well overall, subtle risk variations between adjacent categories remain a challenge. Despite these misclassification trends, the model maintains strong overall performance, particularly in handling the majority classes. To better understand the factors driving model predictions, feature importance scores were extracted from the best-performing model, XGBoost. The analysis revealed that 'OD Days', 'Product Line' and 'Disbursed Amounts' were the top contributors to the model's predictions. These findings enhance the interpretability of the model by aligning with established financial theory and practical lending practices, thereby reinforcing the model's credibility.

B. Evaluating the Impact of Using the Oversampling Technique SMOTE on ML Algorithms to Address Class Imbalance

The second experiment aimed to reduce misclassification and overfitting by using the SMOTE oversampling technique to address the dataset's imbalance. Imbalanced datasets can lead to incorrect predictions, as models may favor the majority class present in the training data. To make the model more generalizable, this experiment increased instances of minority classes, aiming to produce a more robust model. Table 5.2 illustrates the class distribution in the training data after applying SMOTE.

Table 5.2: Training Counts After SMOTE Oversampling

Risk Class	**Post-SMOTE Count**
DOUBTFUL	5000
LOSS	2800
NORMAL	7246
SUBSTANDARD	3000
WATCH	3600

SMOTE was applied with a sampling strategy designed to approximately double the instances of each minority class, instead of forcing all classes to be equal, which can lead to overfitting. This approach to partially increase the minority classes aimed to enhance the model's ability to learn from underrepresented classes by increasing their representation in the training data without overfitting to any particular class. To illustrate the impact of SMOTE on class distribution, fig. 5.7 compares the class frequencies before and after oversampling. This visualization highlights how the minority classes were increased while maintaining a balanced yet realistic distribution, ensuring the model gains better representation without excessive synthetic data generation.

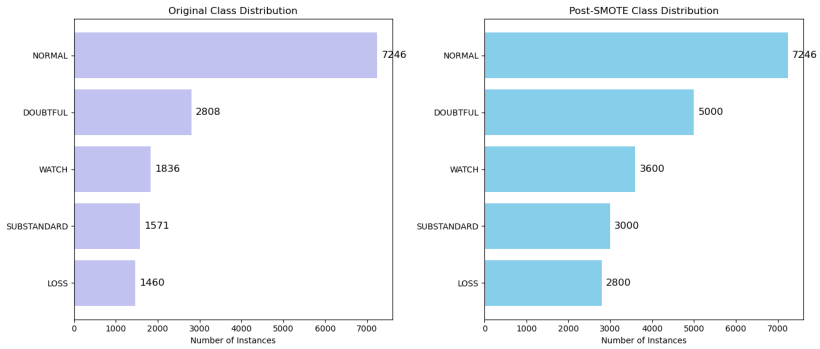


Figure 5.7: Original vs Post SMOTE Class Distribution

Post-SMOTE, none of the classes comprised less than 10% of the total dataset, and the majority class was not more than three times the size of the smallest class. This indicates that the dataset successfully met the defined threshold for a balanced distribution, ensuring a more equitable representation of all classes for improved model performance. After applying SMOTE to the training data, there were improvements to the performance metrics suggesting that the more balanced dataset allowed the classifiers to better learn the decision boundaries of each class, leading to more accurate and reliable predictions. The performance metrics were as shown in table 5.3.

Table 5.3: Performance Metrics for Each Model after SMOTE

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.9368	0.9316	0.9325	0.9320
Random Forest	0.9575	0.9594	0.9495	0.9541
LightGBM	0.9565	0.9581	0.9485	0.9530
XGBoost	0.9552	0.9567	0.9470	0.9515

Random Forest emerges as the best-performing model in this setting, achieving the highest accuracy (0.9575), precision (0.9594), recall (0.9495), and F1-score (0.9541). This suggests that Random Forest effectively leverages the synthetic minority samples to maintain a strong balance between precision and recall. XGBoost, while still competitive, slightly lags behind Random Forest in all metrics, suggesting that it may not gain as much from SMOTE as it did in the imbalanced scenario.

All models demonstrate improved recall scores compared to the imbalanced dataset, confirming that oversampling helps in addressing class imbalance by reducing bias towards the majority class. While recall improves across all models, precision also increases, showing that SMOTE does not lead to excessive false positives. The Random Forest, LightGBM, and XGBoost models show slight improvements in F1-score, with Random Forest achieving the highest at

0.9541. Overall, SMOTE enhances model performance by making predictions more balanced as shown in fig. 5.8.

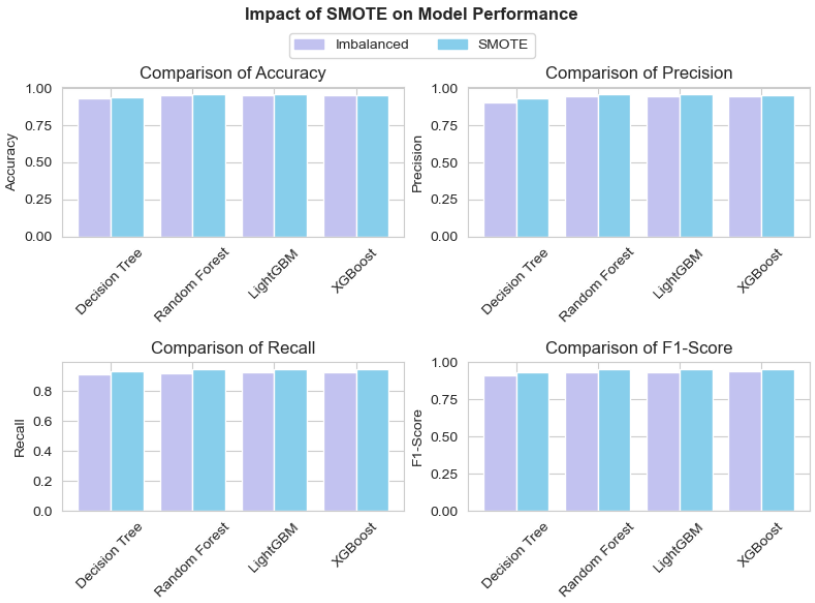


Figure 5.8: Impact of SMOTE on Model Performance

5.4 Conclusion

This chapter outlined the results of the two key aspects that were investigated in this paper: a comparative analysis of multiple machine learning algorithms and the impact of SMOTE on imbalanced data. The first experiment revealed that ensemble-based methods such as Random Forest, LightGBM, and XGBoost consistently outperformed traditional decision trees, demonstrating superior predictive accuracy and overall performance. Among them, XGBoost emerged as the best-performing model across all evaluation metrics on the original dataset. This can be attributed to several of XGBoost's strengths. First, its built-in mechanisms and advanced regularization techniques (L1 and L2) help prevent overfitting, which is critical in real-world financial modeling where generalization is essential. Second, its boosting framework allows it to build trees sequentially, enabling each new tree to correct the errors of the previous ones and thus refine predictions more effectively. Third, XGBoost has demonstrated superior performance on imbalanced datasets, as it can be finely tuned to focus on minority class prediction without compromising overall accuracy. The second experiment explored the role of SMOTE in addressing class imbalance. The results confirmed that SMOTE improved model performance while ensuring that minority classes are better represented in predictions. Specifically, the improvement in recall suggests that SMOTE effectively reduces bias toward the majority class.

SMOTE was applied with a sampling strategy designed to approximately double the instances of each minority class. This approach proved effective in improving model generalization, as it increased the representation of underrepresented classes without leading to significant overfitting. While SMOTE led to a slight performance boost across models, the results show that Random Forest slightly outperformed XGBoost after applying SMOTE, but the difference was marginal. Given these findings, XGBoost remains the most robust model overall. While the performance gains from SMOTE were modest, it proved effective as a preprocessing step to enhance class balance. These results highlight the importance of selecting both the appropriate machine learning algorithm and preprocessing techniques to achieve balanced and reliable classification performance.

6. Conclusion, Recommendations and Future Work

6.1 Conclusion

Managing credit risk effectively is essential for the financial stability, profitability, and growth of lending institutions. This research aimed to improve credit risk assessment within Kenyan lending institutions by evaluating multiple machine learning algorithms and investigating the impact of oversampling techniques on imbalanced data. Specifically, the study sought to identify the most effective approach for multi-class loan classification and develop a predictive model that enables lenders to make data-driven decisions, mitigate risks, and manage loan portfolios more efficiently. The findings of the study reveal that ensemble-based methods, including Random Forest and XGBoost, consistently outperformed traditional decision trees, demonstrating superior predictive accuracy and overall performance. The use of ensemble learning techniques proved effective in improving classification performance, with XGBoost emerging as the best-performing model across all evaluation metrics highlighting its effectiveness in multi-class loan classification. As the dataset was heavily skewed toward normal-risk loans, the study also aimed to evaluate the impact of SMOTE on handling class imbalance in multi-class loan classification. The results showed that applying oversampling techniques improved model performance, particularly in terms of recall. This enhancement was observed across all classifiers, indicating that SMOTE effectively ensured better representation of minority risk categories and reduced model bias toward the majority class. The findings of this study underscore the effectiveness of ensemble-based machine learning models, particularly XGBoost, in enhancing the accuracy of multi-class loan classification. The evaluation of SMOTE further reinforced its importance in addressing class imbalance in loan datasets, demonstrating that oversampling techniques significantly improved recall for minority risk categories. These insights hold significant implications for financial institutions seeking to refine their credit risk assessment and loan portfolio management processes. By leveraging XGBoost and SMOTE, lenders can develop more reliable predictive models leading to more informed lending decisions, better loan portfolio management, and ultimately reduced financial losses.

6.2 Recommendations

Based on the findings of this study and the successful deployment of the predictive model, several key recommendations can be made for financial institutions seeking to

improve their credit risk assessment and loan portfolio management. The predictive model developed in this research can be integrated into existing loan management workflows to enhance risk assessment capabilities and enable lenders to make more data-driven decisions. The tool can be incorporated into automated decision-making systems, assisting loan officers by predicting loan categories, monitoring movement between risk categories to mitigate risk, and tailoring solutions to help curb defaults. To maintain model effectiveness, regular hyperparameter tuning and retraining with updated data are recommended. Additionally, improving data quality and consistency through standardized reporting, particularly in handling missing values, can further enhance the model's reliability.

6.3 Future Work

While this research has demonstrated the effectiveness of tree-based models and the impact of oversampling techniques in multi-class loan classification, there are several areas for further improvement. Testing more advanced algorithms, such as deep learning models, including neural networks, could provide insights into their effectiveness in capturing complex relationships within loan classification data. Additionally, while SMOTE improved classification performance, exploring more advanced versions of SMOTE (e.g., Borderline-SMOTE or ADASYN) may further enhance the representation of minority classes while minimizing the introduction of synthetic noise. Varying the sampling strategy for SMOTE could help enhance model performance improvement while mitigating the risk of overfitting. Beyond oversampling, alternative methods for addressing data imbalance,, could be investigated to enhance model robustness and improve classification performance across all loan categories when class imbalance cannot be solved by resampling methods like SMOTE.

References

Alam, M. (2023). Random forest algorithm: An in-depth guide for data science enthusiasts. Accessed: 2025-02-17.

Araka, H. (2022). *Effect of Loan Portfolio Management, Interest Rate Regulation and Financial Performance of Commercial Banks in Kenya*. PhD thesis, JOOUST.

Aris, A. S. and Rahimi, E. (2023). The impact of loan portfolio management on credit risk: Evidence from banking sector of afghanistan. *Journal of Economics, Finance and Accounting Studies*, 5(5):12–22.

Atellu, A. R. (2021). *Inclusive Finance, Bank Regulation, Concentration and Financial Stability in Kenya*. PhD thesis, University of Nairobi.

Bao, W., Lianju, N., and Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128:301–315.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*.

Berriri, M., Djema, S., Rey, G., and Dartigues-Pallez, C. (2021). Multi-class assessment based on random forests. *Education Sciences*, 11(3):92.

Bhatore, S., Mohan, L., and Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, 4(1):111–138.

Bitok, S. K., Cheboi, J., and Kemboi, A. (2020). Does portfolio quality influence financial sustainability? a case of microfinance institutions in kenya. *Journal of Economics and Financial Analysis*, 3(2):23–39.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Brown, K. and Moles, P. (2014). Credit risk management. *K. Brown & P. Moles, Credit Risk Management*, 16.

CBK (2013). Risk management guidelines.

Chaudhary, A., Kolhe, S., and Kamal, R. (2016). An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*, 3(4):215–222.

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

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792.

Danandeh Mehr, A. (2021). Drought classification using gradient boosting decision tree. *Acta Geophysica*, 69(3):909–918.

Del Moral, P., Nowaczyk, S., and Pashami, S. (2022). Why is multiclass classification hard? *IEEE Access*, 10:80448–80462.

Elreedy, D. and Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences*, 505:32–64.

Elreedy, D., Atiya, A. F., and Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Machine Learning*, 113(7):4903–4923.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gao, B. and Balyan, V. (2022). Construction of a financial default risk prediction model based on the lightgbm algorithm. *Journal of Intelligent Systems*, 31(1):767–779.

Gichuki, J. A. W., Njeru, A., and Tirimba, O. I. (2014). Challenges facing micro and small enterprises in accessing credit facilities in kangemi harambee market in nairobi city county, kenya. *International Journal of Scientific and Research Publications*, 4(12):1–25.

Gongera, E. G. D., Miroga, J. B. D., Ngaruiya, N. W., Mindila, R., Mobisa, M. J., Ongeri, J., Mandere, E. N., and Moronge, M. O. D. (2013). An analysis of loan portfolio management on organization profitability: Case of commercial banks in kenya.

Goyal, D., Choudhary, A., Pabla, B., and Dhami, S. (2020). Support vector machines based non-contact fault diagnosis system for bearings. *Journal of Intelligent Manufacturing*, 31:1275–1289.

Guleria, K., Sharma, S., Kumar, S., and Tiwari, S. (2022). Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning. *Measurement: Sensors*, 24:100482.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer.

Hangkawidjaja, A. D., Prijono, A., Suherman, J., et al. (2021). Discrete cosine transform and multi class support vector machines for classification cardiac atrial arrhythmia and cardiac normal. In *Journal of Physics: Conference Series*, volume 1858, page 012093. IOP Publishing.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee.

Itoo, F., Meenakshi, and Singh, S. (2021). Comparison and analysis of logistic regression, naïve bayes and knn machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, 13(4):1503–1511.

Karekaho, S. S. (2009). *Loan portfolio management and performance of micro finance institutions in Uganda: The case of Wakiso District*. PhD thesis, Makerere University.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kingsford, C. and Salzberg, S. L. (2008). What are decision trees? *Nature biotechnology*, 26(9):1011–1013.

Kinyanjui, H. W. (2013). *Relationship between exchange rate fluctuations and the demand for credit among Commercial Banks in Kenya*. PhD thesis, University of Nairobi.

Lavanya, M., Kannan, P. M., and Arivalagan, M. (2021). Lung cancer diagnosis and staging using firefly algorithm fuzzy c-means segmentation and support vector machine classification of lung nodules. *International Journal of Biomedical Engineering and Technology*, 37(2):185–200.

Le, T.-T.-H., Oktian, Y. E., and Kim, H. (2022). Xgboost for imbalanced multiclass classification-based industrial internet of things intrusion detection systems. *Sustainability*, 14(14):8707.

LightGBM (2025). Lightgbm 4.6.0.99 documentation.

Liu, W., Fan, H., and Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, 189:116034.

Lumumba, V. W., Kiprotich, D., Mpaine, M. L., Makena, N. G., and Kavita, M. D. (2024). Comparative analysis of cross-validation techniques: Loocv, k-folds cross-validation, and repeated k-folds cross-validation in machine learning models. *American Journal of Theoretical and Applied Statistics*, 13(5):175–180.

Luvuma, S. (2021). *Loan portfolio management and financial performance of microfinance institutions in Uganda: a case study of Brac Uganda Microfinance Limited Head office, Kampala.* PhD thesis, University of Kisubi.

Madaan, M., Kumar, A., Keshri, C., Jain, R., and Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In *IOP conference series: materials science and engineering*, volume 1022, page 012042. IOP Publishing.

Makena, P., Kubaison, S. T., and Njati, C. I. (2014). Challenges facing women entrepreneurs in accessing business finance in kenya: Case of ruiru township, kiambu county. *Journal of Business and Management*, 16(4):83–91.

Meyer, D. and Wien, F. (2001). Support vector machines. *R News*, 1(3):23–26.

Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., and Lin, S. F. (2021). Explainable ai in credit risk management. *arXiv preprint arXiv:2103.00949*.

Mongina, J., Ouma, D., and Otsyulah, J. (2022). Control activities and credit risk in registered deposit taking saccos in western kenya.

Neelakandan, S. and Paulraj, D. (2020). A gradient boosted decision tree-based sentiment classification of twitter data. *International Journal of Wavelets, Multiresolution and Information Processing*, 18(04):2050027.

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Noor, F. A. (2020). Effect of covid-19 on loan repayment of small businesses in kenya: A case study of eastleigh business community. *European Journal of Business and Strategic Management*, 5(2):1–14.

Odhiambo, L. A. (2013). The effect of changes in interest rates on the demand for credit and loan repayments by small and medium enterprises in kenya. *University of Nairobi*.

Oduor, S. O. (2021). Impact of mobile loan credit during the covid-19 pandemic in kenya.

Okero, E. O. and Waweru, F. W. (2023). Credit risk assessment and loan repayment among development financial institutions. a case of kenya industrial estates limited. *International Journal of Finance and Accounting*, 2(1):21–29.

Oluwagbenga, O. R. (2023). Navigating data chaos: The power of crisp-dm framework. Accessed: 2025-02-17.

Ong, C.-S., Huang, J.-J., and Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert systems with applications*, 29(1):41–47.

Paudel, S. B., Devkota, B., and Timilsina, S. (2023a). Multi-class credit risk analysis using deep learning. *Journal of Engineering and Sciences*, 2(1):82–87.

Paudel, S. B., Devkota, B., and Timilsina, S. (2023b). Multi-class credit risk analysis using deep learning. *Journal of Engineering and Sciences*, 2(1):82–87.

Pisner, D. A. and Schnyer, D. M. (2020). Support vector machine. In *Machine learning*, pages 101–121. Elsevier.

Ponsam, J. G., Gracia, S. J. B., Geetha, G., Karpaselvi, S., and Nimala, K. (2021). Credit risk analysis using lightgbm and a comparative study of popular algorithms. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, pages 634–641. IEEE.

Raghuwanshi, B. S. and Shukla, S. (2020). Smote based class-specific extreme learning machine for imbalanced learning. *Knowledge-Based Systems*, 187:104814.

Rehman, Z. U., Muhammad, N., Sarwar, B., and Raz, M. A. (2019). Impact of risk management strategies on the credit risk faced by commercial banks of balochistan. *Financial Innovation*, 5:1–13.

ReliefWeb (2023). Kenya economic update, december 2023 (edition 28).

Rokach, L. and Maimon, O. (2005). Decision trees. *Data mining and knowledge discovery handbook*, pages 165–192.

Saberian, M., Delgado, P., and Raimond, Y. (2019). Gradient boosted decision tree neural network. *arXiv preprint arXiv:1910.09340*.

Sadok, H., Sakka, F., and El Maknouzi, M. E. H. (2022). Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, 10(1):2023262.

Scikit-Learn (2023). *RandomForestClassifier — scikit-learn 1.3.0 documentation*.

Scikit-Learn (2024). *Decision Trees*. Accessed: 2025-02-17.

Seynhaeve, J. A. (2022). *The Ethics of Data Privacy*. Doctoral dissertation, Stellenbosch University, Stellenbosch.

Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170362.

Shanmugasundaram, P. (2023). Machine learning episode 11: Tree-based regression models — decision tree random forest. Accessed: 2025-02-17.

Tian, Z., Xiao, J., Feng, H., and Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. *Procedia Computer Science*, 174:150–160.

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, 63:101413.

Umagba, A. O., Abara, B., Isa, Z., Okoro, E., and Yahaya, M. (2022). A multi class machine learning model for predicting credit default in credit risk management. *Available at SSRN 4105836*.

Vosloo, P. G. and Styger, P. (2009). The process approach to the management of loan portfolios. *Journal of Economic and Financial Sciences*, 3(2):171–188.

Wamalwa, N. and Jagongo, A. (2017). Loan portfolio management and firm performance: Theoretical paper review. *International Journal of Management and Commerce Innovations*, 5(2):638–643.

Wang, J., Rong, W., Zhang, Z., and Mei, D. (2022). Credit debt default risk assessment based on the XGBoost algorithm: An empirical study from china. *Wireless Communications and Mobile Computing*, 2022:8005493:1–8005493:14.

Wang, T., Liu, R., Liu, J., and Qi, G. (2024). A novel ensemble model of multiclass credit assessment based on multi-source fusion theory. *Journal of Intelligent & Fuzzy Systems*, 46(1):419–431.

Webb, G. I., Keogh, E., and Miikkulainen, R. (2010). Naïve bayes. *Encyclopedia of machine learning*, 15(1):713–714.

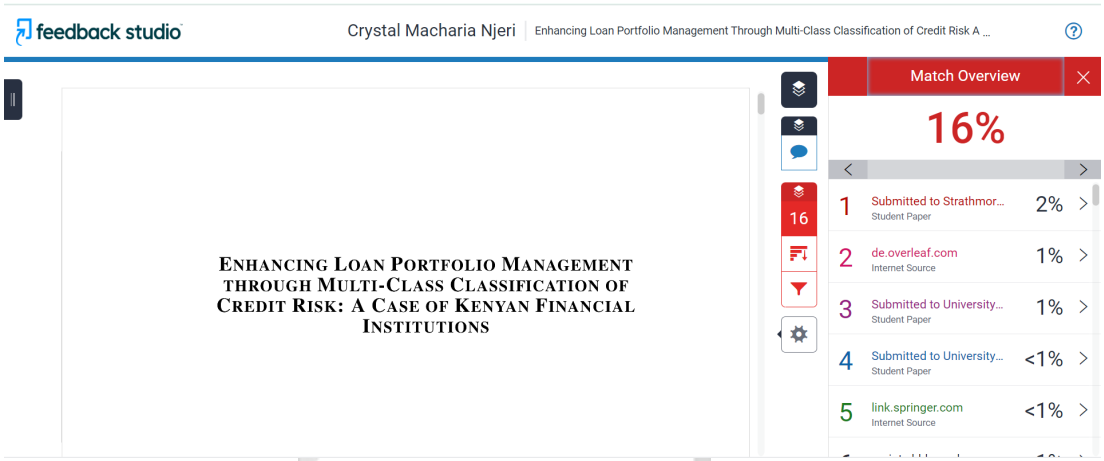
Wei, G., Mu, W., Song, Y., and Dou, J. (2022). An improved and random synthetic minority oversampling technique for imbalanced data. *Knowledge-based systems*, 248:108839.

Wickramasinghe, I. and Kalutarage, H. (2021). Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, 25(3):2277–2293.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.

Zhang, S., Hu, Y., and Tan, Z. (2019). Research on borrower's credit classification of p2p network loan based on lightgbm algorithm. *International Journal of Embedded Systems*, 11(5):602–612.

Appendix A: Similarity Report



Appendix B: Ethical Clearance Release Letter

Strathmore UNIVERSITY

6th June 2024

Crystal Njeri Macharia
078624
crystal.macharia@strathmore.edu

Dear Crystal,

RE: Enhancing Loan Portfolio Management through Multi-Class Classification of Credit Risk: A Case of Kenyan Financial Institutions

This is to inform you that the Office of Graduate Studies on 6th June 2024 received your acknowledgement of breach in ethical processes given that you have already collected data and proceeded to write the Thesis prior to obtaining Ethical clearance. The ethics approval process is ONLY done before any collection of primary or secondary data.

This is a letter for you to proceed with the next steps of your academic requirements.

Please be advised, that in future, all research proposals should be submitted to the SU-ISERC through the RHInnO Ethics platform: https://strathmoreuniversity.rhinno.net/login

Disclaimer: *1) This is not in any way an ethical approval letter. 2) Should there be any legal implications/actions emanating from the research in terms of any ethical violations, you will be personally liable.*

Yours sincerely,

Dr. Bernard Shibwabo
Director of Graduate Studies

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu