Enhancing Credit Scoring in Emerging Markets: Overcoming Data Scarcity with Advanced Machine Learning and Data Augmentation Techniques

By

Regina Wanjiru Gathimba

169122

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

Institute of Mathematical Sciences and @iLabAfrica
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.

Student's Name: **Regina Wanjiru Gathimba**

Sign: _____ Date: 05/22/2025

Approval

The dissertation of **Regina Wanjiru Gathimba** was reviewed and approved by the following:

Dr. John Olukuru

Head of Data Science and Analytics

Strathmore University.

Dr. Godfrey Madigu

Dean, Institute of Mathematical Sciences

Strathmore University.

Prof. Bernard Shibwabo

Director of Graduate Studies

Strathmore University.

Abstract

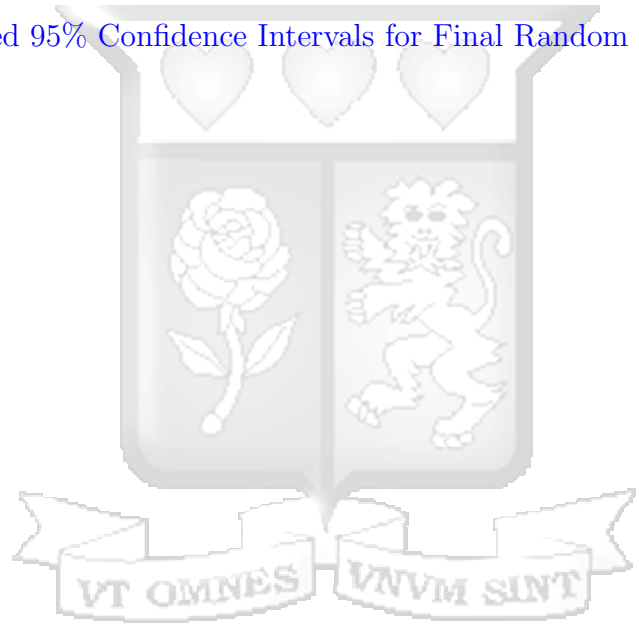
Credit risk assessment is essential for lending institutions, especially in data-scarce environments where limited borrower information complicates accurate risk evaluation. This study presents a robust machine learning pipeline that integrates real demographic data with synthetic financial records generated via a Conditional Tabular GAN (CTGAN) model, effectively augmenting the training dataset. Exploratory Data Analysis (EDA) revealed that debt-to-income and debt-to-savings ratios were among the most predictive features; these were log-transformed to address skewness and improve model learning. Four classification models Logistic Regression, Random Forest, Gradient Boosting and Neural Network were trained and evaluated. The Random Forest model consistently outperformed others when trained on a 75% real / 25% synthetic mixed dataset, achieving an accuracy of 75%, a macro F1-score of 0.69, and an AUC-ROC of 68.6%. To improve statistical reliability, bootstrapped confidence intervals were computed, confirming model robustness. A fairness analysis was also conducted by excluding sensitive attributes such as sex and marital status, resulting in an ethically aligned model without significant performance loss. The final Random Forest model was deployed using a Streamlit web application, enabling real-time credit scoring via a lightweight and user-friendly interface. This research demonstrates that synthetic data augmentation, combined with advanced machine learning, can enhance credit scoring in emerging markets, particularly for microfinance institutions. Future work will focus on fairness auditing, model calibration, and integration into financial infrastructure to maximize operational impact.

Key Words: credit scoring, Random Forest, CTGAN, synthetic data, emerging markets, fairness, microfinance

Table of Contents

List of Figures

List of Tables



List of Abbreviations

API Application Programming Interface

AUC Area Under Curve

CRISP-DM Cross-Industry Standard Process for Data Mining

CSV Comma Separated Values

EDA Exploratory Data Analysis

GBM Gradient Boosting Machine

HTML Hypertext Markup Language

ICT Information and Communications Technology

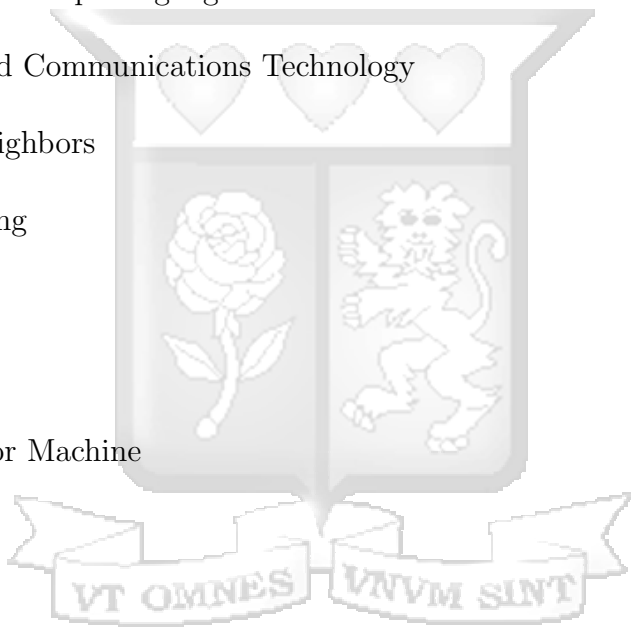
KNN K-Nearest Neighbors

ML Machine Learning

NN Neural Network

RF Random Forest

SVM Support Vector Machine



Acknowledgments

I am sincerely thankful to all those who have supported and guided me throughout the journey of completing this dissertation. Above all, I thank the Almighty God for His endless grace, strength, and guidance that have seen me through every stage of this academic endeavor.

I extend my profound appreciation to my supervisor, Dr. John Olukuru, for his unwavering support, insightful feedback, and valuable mentorship that have significantly influenced the direction and quality of this research. His encouragement and expertise have been instrumental in shaping my academic growth.

I am also grateful to my fellow classmates for their constant encouragement, collaboration, and companionship throughout the course. Special thanks go to the university's administrative and academic staff, as well as the wider Strathmore University community, for providing the necessary resources and creating an environment conducive to learning and research.

Lastly, I acknowledge the work of the scholars and researchers whose literature formed the foundation of my study. Their dedication to advancing knowledge has been a vital source of inspiration and insight throughout this research process.

Chapter 1: Introduction

1.1 Background

1.1.1 Global Context

The global financial system depends on accurate credit assessments to evaluate the creditworthiness of individuals and businesses. Credit scoring models enable financial institutions to assess risk effectively, facilitating access to credit and promoting economic growth. In developed economies, these models benefit from structured and comprehensive financial data, including detailed loan repayment histories and credit utilization patterns, allowing for precise risk evaluations Sahay et al. (2015). The availability of such data supports well-functioning credit markets and fosters financial inclusion.

Despite advancements in financial services, financial exclusion remains a persistent challenge worldwide. The 2021 Global Findex Database estimates that approximately 1.4 billion adults remain unbanked, with the majority residing in developing regions Demirgüç-Kunt et al. (2020). The primary barrier to formal financial access in these economies is the lack of structured financial histories, which prevents individuals from securing loans for essential investments such as education, healthcare, and entrepreneurship. In contrast, economies with well-established financial infrastructures and centralized credit systems experience significantly lower credit exclusion rates, enabling more inclusive financial growth Beck et al. (2009).

Emerging markets, particularly in developing economies, face significant barriers to financial deepening due to data inefficiency, weak credit reporting frameworks, and inconsistent financial records Wagdi and Tarek (2022). These challenges create a cycle where limited financial data restricts access to credit, and a lack of credit history further inhibits the development of structured financial data Sahay et al. (2015). Research highlights the need for credit scoring models that can adapt to data-scarce environments, emphasizing the importance of incorporating borrower-specific characteristics to improve risk assessment despite limited historical data Hasan (2016).

As financial institutions seek to expand credit access in underserved regions, recent studies explore alternative approaches to improving credit risk assessment in data-limited settings Ampountolas et al. (2021). Understanding these challenges is essential to addressing

financial exclusion and strengthening credit systems, particularly in emerging markets where traditional scoring models may be ineffective.

1.1.2 Regional Context

Sub-Saharan Africa's financial landscape remains challenging, marked by low financial inclusion, weak financial infrastructure, and the absence of comprehensive credit reporting systems. According to the 2021 Global Findex Database, approximately 45% of adults in the region remain unbanked, significantly limiting their access to formal financial services Demirgüç-Kunt et al. (2020). The lack of structured credit histories particularly affects micro, small, and medium enterprises (MSMEs), which play a critical role in the region's economy. Without access to affordable credit, these businesses struggle to expand, hindering job creation and economic growth Ampountolas et al. (2021).

Traditional credit scoring models, which rely on structured financial data such as income records and loan repayment histories, are often ineffective in Sub-Saharan Africa due to fragmented and inconsistent data Lessmann et al. (2015). Many financial institutions lack reliable borrower information because of cash-based transactions, weak credit infrastructures, and the absence of centralized credit bureaus. This forces lenders to either take on higher risks or exclude large segments of the population from credit access. In turn, individuals and MSMEs resort to high-cost informal lending systems, which limit financial growth opportunities Wagdi and Tarek (2022).

Beyond data limitations, institutional and regulatory weaknesses further restrict credit accessibility. Many countries in the region lack clear policies for credit reporting and data sharing, making it difficult to establish standardized borrower profiles. Financial institutions are often forced to rely on manual risk assessment processes, which are subjective and inconsistent Atieno (2009). These inefficiencies discourage formal lenders from extending credit to underserved populations, perpetuating financial exclusion.

Given these challenges, recent research explores alternative credit assessment approaches that can function in data-limited environments. Studies indicate that machine learning-based models can analyze borrower behavior patterns and improve credit risk evaluations where traditional models fail Lessmann et al. (2015). While these approaches remain in early research stages, they highlight the potential for innovation in expanding credit

access in Africa Wagdi and Tarek (2022).

1.1.3 Local Context

Kenya is often regarded as a financial inclusion success story due to the widespread adoption of mobile banking platforms such as M-Pesa, which has expanded access to financial services across the country. According to the 2021 Global Findex Database, 79% of adults in Kenya have a mobile money account, significantly higher than the Sub-Saharan Africa average Demirgüç-Kunt et al. (2020). While this advancement has improved access to basic financial services, significant barriers remain in credit accessibility, particularly for rural populations and small businesses. The absence of structured credit histories and the reliance on informal financial mechanisms continue to limit access to loans from formal financial institutions Atieno (2009). These challenges contribute to persistent economic disparities between urban and rural areas, where access to financial resources is unevenly distributed.

Small and medium enterprises (SMEs), which form the backbone of Kenya's economy, face similar obstacles in accessing credit. SMEs account for approximately 80% of employment and contribute over 30% to the country's GDP, yet they remain underserved by formal lenders due to the lack of financial records and weak credit reporting systems Bowen et al. (2009). Many businesses operate in the informal sector, where cash transactions dominate, making it difficult to establish creditworthiness using traditional scoring models. Additionally, regulatory and institutional barriers, including high collateral requirements and limited credit information sharing, further restrict SME financing opportunities Wagdi and Tarek (2022). As a result, many enterprises turn to informal credit sources, such as community savings groups and private lenders, which often charge high-interest rates and provide inadequate loan amounts to support business expansion.

The financial system in Kenya has made progress in developing credit reporting mechanisms, with Credit Reference Bureaus (CRBs) introduced in 2010 to enhance risk assessment for lenders. However, the effectiveness of these systems is limited, as they primarily cover borrowers with formal financial records while excluding those in informal and rural economies Lessmann et al. (2015). This gap in financial infrastructure highlights the need for research on alternative credit assessment methods that can operate in data-

scarce environments. Recent studies explore the potential for advanced credit risk models that account for borrower behavior and transaction patterns, offering new insights into addressing credit exclusion Ampountolas et al. (2021).

1.2 Problem Statement

Financial institutions in emerging markets, particularly microfinance institutions and banks, face significant challenges in accurately assessing credit risk due to limited and unreliable financial data. The absence of comprehensive credit histories and structured financial records leads to inconsistent risk assessments, making it difficult for lenders to determine borrower creditworthiness. As a result, large segments of the population, including individuals and small businesses, remain excluded from formal financial services. This financial exclusion restricts access to affordable credit, limiting entrepreneurial growth, investment opportunities, and overall economic development in these regions. Despite advancements in data-driven credit assessment, traditional credit scoring models remain inadequate in data-scarce environments, necessitating further research into alternative risk evaluation techniques tailored for emerging markets.

1.3 Research Objectives

1.3.1 Main Objective

To critically evaluate the effectiveness and limitations of traditional credit scoring models in data-scarce environments, particularly in emerging markets, by assessing their predictive performance and reliability using key performance metrics.

1.3.2 Specific Objective

I. To develop and optimize machine learning models incorporating data augmentation techniques, such as Generative Adversarial Networks (GANs), for improved credit risk assessment in data-scarce environments.

II. To systematically evaluate and validate the impact of data augmentation techniques on accuracy, fairness, and generalizability in credit scoring models, using predefined performance benchmarks.

III. To provide policy recommendations for microfinance institutions and regulators,

based on findings showing the benefits of synthetic data augmentation, fairness-aware modeling, and model deployment feasibility, in order to improve equitable credit access in emerging markets.

1.4 Research Questions

I. How do GAN-augmented machine learning models improve credit scoring in data-scarce environments?

II. How effective are GANs and other data augmentation techniques in improving the accuracy, fairness, and adaptability of credit scoring models?

III. What policy recommendations can be derived from the research findings to improve credit accessibility in emerging markets?

1.5 Scope and Limitations of the Study

1.5.1 Scope

This study focuses on enhancing credit risk assessment in data-scarce emerging markets, with a primary geographical focus on Sub-Saharan Africa. The research evaluates the effectiveness of machine learning models, particularly ensemble methods and optimized data augmentation techniques, in addressing data scarcity challenges. While the findings are expected to be applicable to other emerging markets with similar credit data constraints, they may not directly generalize to regions with well-established credit reporting infrastructures.

The study excludes alternative data sources, such as mobile transaction histories and social media data, focusing instead on demographic data and synthetic financial data to improve credit scoring. Additionally, the research does not analyze regulatory or institutional frameworks governing credit scoring systems. However, policy recommendations will be provided to ensure the ethical and fair use of synthetic data in credit risk assessment.

1.5.2 Limitations

The study's scalability and generalizability may be constrained by variations in demographic data availability across different emerging markets. While synthetic data tech-

niques will be used to simulate data scarcity, the findings may not fully replicate real-world financial environments, where economic conditions, credit structures, and borrower behaviors vary significantly.

Additionally, the study does not address the regulatory implications of synthetic data in financial decision-making, meaning that real-world deployment may require further policy and legal considerations. While the machine learning models will be optimized for accuracy and fairness, their performance may require customization based on specific regional and economic factors. Finally, data privacy and ethical concerns surrounding synthetic data usage remain an open challenge that must be further investigated before large-scale adoption.

1.6 Research Justification

Limited financial histories in emerging markets hinder access to credit, restricting economic participation and financial inclusion. This study explores how data augmentation and machine learning improve credit risk assessment in data-scarce environments.

1.6.1 Financial Institutions

Lenders struggle to assess creditworthiness, leading to high-interest rates or exclusion. This study enhances risk assessment through AI-driven models, reducing defaults and expanding financial access.

1.6.2 Policymakers and Regulators

Existing regulations lack provisions for synthetic data in credit scoring. This research provides insights to inform ethical and fair policy frameworks.

1.6.3 Economic and Social Impact

Better credit access supports SMEs, job creation, and poverty reduction. By improving risk assessment, this study promotes financial inclusion and economic resilience.

Chapter 2: Literature Review

2.1 Theoretical Framework

2.1.1 Information Asymmetry and Credit Scoring

The theory of information asymmetry, introduced by Akerlof (1970), is foundational to understanding credit market inefficiencies in emerging economies. Information asymmetry occurs when borrowers possess more financial knowledge than lenders, leading to adverse selection where high-risk borrowers secure loans at the same rates as low-risk borrowers and moral hazard, where borrowers default due to weak monitoring Akerlof (1978).

In developed economies, financial institutions mitigate information asymmetry by leveraging structured and comprehensive financial data, such as centralized credit bureau reports, credit repayment histories, and borrower income records Demirgüç-Kunt et al. (2020). These datasets enable precise credit risk assessments, reducing default rates and increasing financial stability . However, in emerging markets, financial records are often fragmented, unavailable, or informal, making risk assessment difficult and increasing financial exclusion .

The consequences of information asymmetry are particularly severe in regions with low financial infrastructure, where lenders lack access to verified borrower data. Demirgüç-Kunt et al. (2020) estimate that nearly 1.4 billion adults worldwide remain unbanked, primarily due to the absence of formal credit histories and structured financial records. As a result, traditional credit scoring models often fail to assess borrower risk accurately, restricting access to formal credit and increasing reliance on high-cost informal lending systems.

Recent advancements in machine learning-based credit scoring models provide potential solutions to mitigate information asymmetry by identifying borrower characteristics beyond traditional financial metrics. These models analyze borrower behavior patterns, such as repayment consistency and transaction histories, to generate more accurate risk assessments Suhadolnik et al. (2023). However, the deployment of machine learning in credit scoring raises ethical concerns, including bias, fairness, and model interpretability, which must be addressed for responsible adoption Frost et al. (2019).

2.1.2 Risk Management and Machine Learning in Credit Scoring

Effective risk management is essential for credit assessment, ensuring that lenders can distinguish creditworthy borrowers from high-risk applicants. Traditional statistical credit risk models, such as the Altman Z-score Altman (1968) and the Merton Model Merton (1974), estimate default probabilities based on structured financial data. These models rely on balance sheets, income statements, and credit repayment histories, making them highly effective in data-rich environments but less reliable in data-scarce settings Louzada et al. (2016).

In emerging markets, where financial data is often fragmented or outdated, traditional credit scoring models struggle to generate consistent risk assessments, often leading to financial exclusion Suhadolnik et al. (2023). To address these limitations, financial institutions are increasingly adopting machine learning (ML) models, which can process unstructured and incomplete datasets while improving predictive accuracy Mhlanga (2021).

Machine learning techniques such as Random Forests, Gradient Boosting Machines (GBM), and Neural Networks offer several advantages in credit risk management:

I. Handling non-linearity in borrower behavior Mhlanga (2021).

II. Identifying hidden patterns in fragmented financial data Suhadolnik et al. (2023).

III. Enhancing predictive accuracy through automated feature selection.

A particularly notable innovation in ML-driven credit scoring is the integration of Generative Adversarial Networks (GANs), which generate synthetic borrower profiles to supplement limited datasets. Singh et al. (2025) demonstrate that GAN-generated synthetic data enhances the robustness of credit scoring models by increasing dataset diversity and reducing imbalances. Similarly, Ramzan et al. (2024) show that GAN-augmented financial datasets improve the accuracy of default risk predictions, particularly in data-constrained environments.

The increasing adoption of ML-based risk models signals a shift away from purely statistical approaches, allowing for more inclusive and adaptive credit assessment methods.

2.1.3 Financial Inclusion and Economic Growth

Financial inclusion is recognized as a key driver of economic growth, particularly in emerging markets, where access to credit is often restricted due to the absence of structured financial records. Demirguc-Kunt et al. (2018) emphasize that financial inclusion fosters entrepreneurship, investment in education, and poverty reduction by enabling previously unbanked populations to access formal credit services. However, traditional credit scoring models disproportionately exclude individuals and businesses without formal credit histories, exacerbating financial inequality Louzada et al. (2016).

In emerging economies, a large proportion of the population engages in informal financial activities, which are not captured by conventional credit reporting systems. This lack of formalized financial data means that many creditworthy borrowers are denied access to loans, limiting their ability to invest in businesses, housing, or education Noriega et al. (2023). Additionally, micro, small, and medium enterprises (MSMEs), which are vital contributors to economic growth, often struggle to secure funding due to limited financial documentation and weak institutional frameworks Suhadolnik et al. (2023).

Machine learning (ML) has emerged as a transformative tool in promoting financial inclusion by enabling data-driven lending decisions that go beyond traditional credit metrics. Ensemble learning techniques, such as Random Forests and Gradient Boosting Machines (GBMs), have demonstrated superior performance in identifying patterns in borrower behavior, even in cases where financial data is scarce or incomplete Noriega et al. (2023).

An emerging approach to further enhance financial inclusion involves Generative Adversarial Networks (GANs), which generate synthetic borrower profiles to augment credit datasets. Singh et al. (2025) highlight that GANs can generate high-quality artificial borrower data, which helps train ML models on more diverse financial behaviors, thereby improving credit accessibility for individuals lacking traditional financial records. Similarly, Ramzan et al. (2024) demonstrate that GAN-augmented datasets improve credit scoring accuracy, reducing false negative classifications that previously excluded creditworthy borrowers from the financial system.

By integrating machine learning and synthetic data augmentation into credit risk assessment, financial institutions can develop more inclusive lending frameworks, thereby

stimulating economic growth.

2.2 Empirical Framework

2.2.1 Credit Scoring Models in Emerging Markets

Traditional credit scoring models, such as logistic regression and decision trees, have underperformed in emerging markets due to incomplete and imbalanced datasets. These models rely heavily on structured financial data, which is often scarce or unavailable in developing economies. Brown and Mues (2012) highlighted that traditional models struggle with imbalanced datasets, where defaulting loans are significantly fewer than non-defaulting ones, leading to reduced predictive accuracy and inconsistent credit decisions.

Empirical research suggests that machine learning (ML) models, particularly ensemble learning methods, outperform traditional credit scoring techniques in emerging markets. Lessmann et al. (2015) conducted a benchmarking study of 41 classification algorithms, concluding that ensemble methods, such as Random Forests and Gradient Boosting Machines (GBM), significantly outperform traditional methods like logistic regression in handling fragmented and incomplete financial data. These models capture non-linear borrower relationships, improving predictive performance where traditional methods fail.

Similarly, Noriega et al. (2023)conducted a systematic literature review showing that ML-based credit scoring models uncover hidden borrower patterns, which traditional techniques fail to detect. These findings emphasize that ML models facilitate financial inclusion by improving credit assessments for underbanked populations, allowing financial institutions to extend fairer lending opportunities to individuals and small businesses without extensive credit histories.

Despite these advancements, challenges remain. Hurlin et al. (2024) highlight that data scarcity in emerging markets continues to hinder ML adoption in financial institutions, as limited financial records and borrower data reduce model performance. Addressing this issue requires innovative data augmentation techniques to supplement missing data, as discussed in the next section.

2.2.2 Data Augmentation Techniques in Financial Modelling

A persistent challenge in credit risk modeling is data scarcity, particularly in emerging markets where formal financial records are limited. Generative Adversarial Networks (GANs) have emerged as a solution to generate synthetic financial data, closely mimicking real-world borrower information.

Singh et al. (2025) highlight that GAN-generated synthetic data enhances credit risk models by supplementing missing data and improving model robustness. By generating realistic borrower profiles, GANs allow credit scoring models to be trained on more diverse financial behaviors, reducing bias and improving predictive accuracy.

Beyond GANs, alternative data augmentation strategies have been explored to improve credit risk assessment. La Gatta et al. (2025) propose an explainable AI-based augmentation strategy that creates synthetic financial data while preserving key statistical properties of real borrower data. Their empirical findings show that augmented datasets significantly improve model accuracy, particularly in credit scoring applications where real data is limited.

In addition, Esteban et al. (2017) initially demonstrated the use of GANs in generating high-quality synthetic datasets, originally in the medical field. However, their methodology has since been adapted to financial modeling, providing a framework for training credit scoring models with enhanced synthetic datasets. Similarly, Ramzan et al. (2024) evaluated GAN-generated synthetic financial data, focusing on stock market prediction. Their findings indicate that synthetic data generation can address inconsistencies in real datasets, improving model performance in financial applications.

While synthetic data presents new opportunities for credit risk assessment, challenges remain regarding the ethical use of synthetic borrower profiles and the potential biases embedded in GAN-generated datasets Frost et al. (2019). Ensuring regulatory compliance and data privacy protections is critical for the widespread adoption of GAN-based credit scoring models in emerging markets.

Limitations and Alternatives to GAN-Based Augmentation

While GANs, particularly CTGAN, have gained traction for generating realistic tabular data, their use comes with notable challenges. GANs are known to suffer from issues

such as training instability, mode collapse, and sensitivity to hyperparameter tuning Goodfellow et al. (2014a). Moreover, they may inadvertently replicate historical biases present in the training data, thereby amplifying systemic unfairness Mehrabi et al. (2021).

Alternative data augmentation methods include Variational Autoencoders (VAEs) Kingma and Welling (2013), SMOTE (Synthetic Minority Over-sampling Technique) Chawla et al. (2002), and traditional bootstrapping. VAEs provide a probabilistic generative framework that can offer more stable convergence during training, while SMOTE generates synthetic samples by interpolating between existing minority class observations. However, these alternatives may not capture complex feature dependencies as effectively as GAN-based models Xu et al. (2019).

CTGAN was selected in this study due to its ability to handle mixed data types and model non-linear feature interactions in high-dimensional tabular data. Nonetheless, it is important to acknowledge its limitations and the need for rigorous fairness evaluation. Future research should explore hybrid augmentation strategies and comparative benchmarking to ensure robustness and ethical alignment across different use cases.

2.2.3 Machine Learning and Financial Inclusion

Machine learning (ML) has emerged as a transformative tool in promoting financial inclusion, particularly in data-scarce environments. Traditional credit scoring models often exclude borrowers without formal financial histories, making it difficult for underbanked populations to access credit. Demirgüç-Kunt et al. (2018) emphasize that financial inclusion fosters economic growth, allowing individuals to invest in businesses, education, and healthcare. However, conventional risk models disproportionately exclude those without documented financial activity, limiting their ability to participate in formal financial systems.

Noriega et al. (2023) highlight that ML-based credit scoring models improve financial accessibility, particularly through ensemble learning techniques, such as Random Forests and GBMs, which have demonstrated superior performance in identifying hidden borrower patterns even in cases where structured credit data is unavailable.

Further, Gao et al. (2023) investigated the use of Long Short-Term Memory (LSTM) networks to predict credit risk in rural areas, where financial data is often fragmented

or incomplete. Their findings reveal that LSTM models capture borrower repayment behavior effectively, providing a more reliable method for rural credit risk assessment. These results support the broader application of ML-based lending decisions for financial inclusion.

However, algorithmic bias in ML models remains a significant concern. Frost et al. (2019) underscore the importance of regulatory frameworks to ensure that ML-based lending decisions remain ethical and unbiased. Without proper oversight, ML models may perpetuate existing financial disparities by reinforcing biases present in training datasets.

By integrating ML-driven credit assessment techniques, financial institutions can expand access to formal credit, promoting entrepreneurship, economic growth, and poverty reduction. However, addressing the ethical and regulatory challenges of ML-based lending will be essential for ensuring fair and responsible financial inclusion strategies.

2.3 Challenges

Despite the increasing adoption of machine learning (ML) models and data augmentation techniques in credit scoring, several persistent challenges hinder their full potential and large-scale implementation in emerging markets. These challenges span technical, ethical, and regulatory dimensions, creating barriers for financial institutions aiming to leverage these innovations effectively.

2.3.1 Data Scarcity and Quality

One of the primary limitations in credit scoring within emerging markets remains the scarcity and poor quality of financial data. While Generative Adversarial Networks (GANs) and other data augmentation techniques have shown promise in addressing data gaps, these methods are still in the early stages of practical implementation.

Esteban et al. (2017) and Ramzan et al. (2024) demonstrate that GAN-generated synthetic data can expand training datasets, but they caution that data quality remains a critical concern. Poorly generated synthetic data may lead to inaccurate models, potentially replicating biases and weaknesses found in traditional credit scoring approaches. This could reinforce financial exclusion rather than mitigating it.

Moreover, financial data in emerging markets is often fragmented, inconsistent, or sourced from informal financial activities, making it difficult for ML models to fully capture borrower behavior. Brown and Mues (2012) highlight that low-quality, unstructured financial data negatively impacts the predictive accuracy of credit models, further complicating efforts to improve lending decisions in these regions.

2.3.2 Algorithmic Bias and Fairness

The risk of algorithmic bias presents a significant challenge to the equitable adoption of ML-based credit scoring models. Frost et al. (2019) and Gao et al. (2023) emphasize that ML algorithms, when trained on incomplete or biased data, can exacerbate existing financial inequalities.

This is particularly problematic in emerging markets, where certain demographic groups (e.g., rural populations, women, and low-income borrowers) may be systematically excluded from formal credit access due to biased algorithms. If ML models primarily learn from existing financial data, they may replicate and reinforce historical lending disparities, thereby reducing the effectiveness of ML-based solutions in promoting financial inclusion.

Additionally, the lack of model interpretability often referred to as the "black box" nature of advanced ML algorithms such as neural networks compounds the issue. Financial institutions require explainable and transparent models to ensure fair lending decisions. Kumar et al. (2021) stress the importance of developing interpretable ML models that can justify their credit risk predictions, particularly in regions where financial exclusion is already a significant issue.

2.3.3 Regulatory and Ethical Concerns

The integration of ML and synthetic data into credit risk assessment raises critical regulatory and ethical challenges. In many emerging markets, there are limited regulatory frameworks governing the use of artificial intelligence (AI) and synthetic data in financial services, creating uncertainty for financial institutions.

Frost et al. (2019) argue that while AI-driven credit scoring has the potential to revolutionize financial inclusion, its implementation must be carefully regulated to ensure

compliance with data privacy, fairness, and transparency standards. A key concern is that synthetic data although designed to protect borrower privacy can still contain identifiable patterns, inadvertently exposing sensitive financial information if not properly managed.

As data protection laws evolve, ensuring compliance with ethical AI standards will be critical for financial institutions adopting GAN-generated synthetic datasets. The development of standardized regulatory frameworks will play a crucial role in facilitating responsible AI adoption in credit scoring.

2.3.4 Integration with Traditional Financial Systems

Beyond regulatory concerns, the technical integration of advanced ML-based credit scoring models with existing financial systems remains a significant challenge. Many financial institutions in emerging markets continue to rely on legacy infrastructure that lacks compatibility with modern AI-based solutions.

Lessmann et al. (2015) highlight that while ML algorithms demonstrate superior predictive accuracy, they require substantial computational power and data storage capacities. Many financial institutions in emerging markets lack the technical expertise and infrastructure necessary to implement and maintain AI-driven credit risk assessment models.

This technological gap makes it difficult for financial institutions to fully capitalize on the benefits of ML and data augmentation, limiting the scalability and widespread adoption of these innovations. Addressing these challenges will require significant investment in digital infrastructure, as well as capacity-building initiatives to equip financial professionals with the necessary technical expertise to manage and deploy AI-driven credit scoring systems.

2.4 Research Gaps and Conceptual Framework

2.4.1 Research Gaps

Despite significant advancements in the application of machine learning models and data augmentation techniques to credit scoring, several key research gaps persist, particularly concerning their practical implementation in emerging markets, where data scarcity remains a major challenge. While studies such as Esteban et al. (2017) and Ramzan et al.

(2024) have demonstrated that Generative Adversarial Networks (GANs) can successfully generate synthetic data to enhance credit scoring models, most existing research has primarily focused on proving the concept rather than assessing long-term performance and real-world applicability. As a result, several critical gaps remain:

1. **Limited Empirical Validation of GAN-Augmented Data in Credit Scoring** While GANs have been shown to improve the quantity of training data, concerns persist regarding the quality and reliability of GAN-generated synthetic data in predicting long-term credit risk. Existing studies primarily emphasize technical feasibility rather than real-world implementation in financial institutions. Specifically, there is limited empirical research evaluating how GAN-augmented datasets affect credit scoring accuracy over time in actual financial settings. Given the high stakes of credit risk assessment, more rigorous testing in applied financial environments is needed to validate the effectiveness of GAN-generated data in improving credit decision-making for underserved populations.

2. **Challenges in Integrating GAN-Generated Data with Traditional Financial Data** Despite their ability to generate synthetic financial records, GANs still face challenges in ensuring that the generated data accurately represents real-world borrower characteristics and risk patterns. One key concern is the potential for GAN-generated data to introduce biases or distortions, which could negatively impact model performance. While Ramzan et al. (2024) highlight the importance of ensuring representativeness in synthetic data, there remains insufficient research on best practices for integrating synthetic and traditional data sources within credit scoring models. This raises critical questions regarding data realism, feature consistency, and the stability of GAN-enhanced models in practical deployment.

3. **Applicability of GAN-Based Credit Scoring in Emerging Markets** Although GAN-based augmentation has been applied in various domains, its use in credit scoring for emerging markets remains underexplored. Many existing studies have focused on developed economies with access to extensive financial datasets, whereas the effectiveness of GAN-generated data in data-scarce environments where financial records are often incomplete or fragmented has not been sufficiently examined. There is a need for further research on how GANs can be effectively trained

using limited financial data from microfinance institutions, rural lenders, and other alternative credit sources to improve credit accessibility for financially excluded populations.

4. **Ethical and Regulatory Considerations in GAN-Augmented Credit Scoring** The increasing reliance on AI-generated data in financial decision-making introduces ethical and regulatory challenges that remain inadequately addressed in existing literature. While some studies, such as Frost et al. (2019), discuss the ethical implications of AI-driven credit scoring, research on the specific risks associated with GAN-generated data such as data privacy concerns, fairness in lending, and potential algorithmic biases is still in its early stages. Furthermore, regulatory frameworks governing synthetic data usage in credit risk assessment are not well defined, particularly in emerging markets where financial regulations are still evolving. This lack of clarity presents a significant barrier to the widespread adoption of GANs in real-world lending environments.

This study aims to fill these research gaps by specifically exploring how traditional financial data in emerging markets can be systematically augmented using GAN-generated synthetic data. By improving dataset quality and model robustness, this research seeks to test the integration of GANs in real-world credit scoring applications, thereby enhancing credit accessibility for underserved populations and contributing to broader financial inclusion.

2.4.2 Conceptual Framework

The conceptual framework for this study is centered on the integration of machine learning algorithms and GAN-generated synthetic data to enhance credit scoring models in data-scarce environments. The framework focuses on expanding traditional financial data through synthetic data generation, addressing the persistent challenge of limited credit data availability in emerging markets. This approach aligns with the theoretical framework, which underscores how machine learning models mitigate information asymmetry and data fragmentation in financial decision-making Akerlof (1970); Demirgüç-Kunt et al. (2020).

The key components of this framework include:

I. **Machine Learning Algorithms for Credit Scoring** Machine learning models such as random forests, gradient boosting machines, and neural networks have demonstrated superior performance in handling incomplete or fragmented datasets. These models can leverage both traditional financial data and GAN-augmented synthetic data to improve credit risk assessment, particularly in contexts where borrower information is scarce. Lessmann et al. (2015) emphasize that ensemble methods outperform traditional credit scoring techniques, making them well-suited for emerging markets with unreliable or missing financial records.

II. **Synthetic Data Augmentation via Generative Adversarial Networks (GANs)** Generative Adversarial Networks (GANs) will be employed to generate synthetic financial data that mimics real-world borrower characteristics. This process enhances dataset richness and addresses gaps in traditional financial records, ultimately improving the predictive performance of credit scoring models. Ramzan et al. (2024) demonstrate that GANs can create high-fidelity synthetic data, filling in missing financial information while minimizing distortions. Additionally, using GAN-generated data offers an alternative to controversial credit scoring methods that rely on social media or mobile transaction data, which raise ethical concerns regarding privacy and fairness.

III. **Enhanced Credit Scoring Models with Augmented Datasets** By integrating machine learning models with GAN-augmented datasets, credit scoring models are expected to become more accurate, reliable, and inclusive. Improved credit risk predictions allow financial institutions to extend credit to previously unbanked populations, reducing financial exclusion. Noriega et al. (2023) highlight that such data-driven enhancements can significantly improve the inclusiveness and fairness of credit scoring models, particularly in data-constrained environments.

IV. **Advancing Financial Inclusion** The ultimate objective of this research is to enable microfinance institutions and other lenders in emerging markets to expand access to credit through more accurate and inclusive credit scoring mechanisms. By improving predictive accuracy and reducing borrower exclusion, this study contributes to greater financial inclusion and economic growth in underserved communities. Gao et al. (2023) demonstrate that machine learning techniques, including

LSTM models and other AI-driven approaches, can be leveraged to enhance financial accessibility in rural and economically marginalized regions.

The conceptual structure guiding this research is summarized in Figure 2.1.

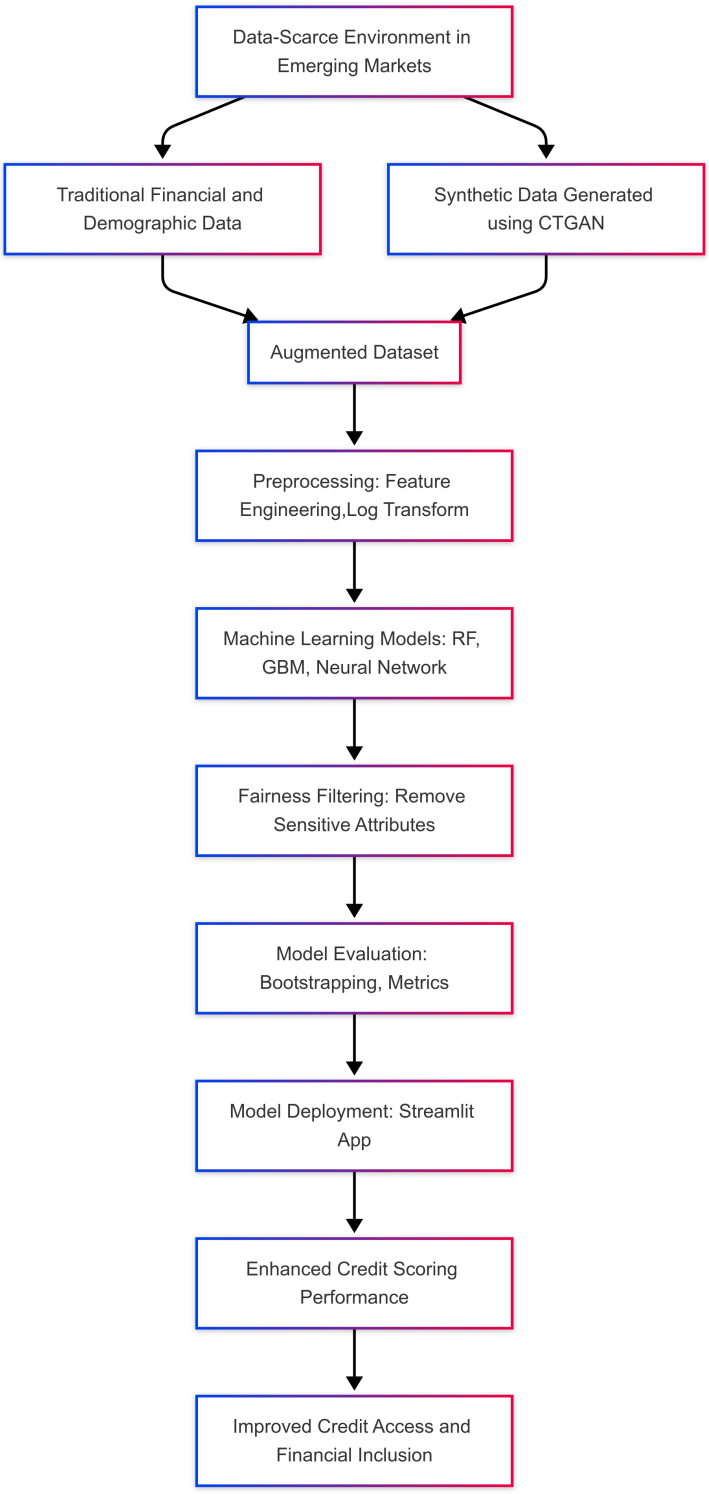


Figure 2.1: Conceptual Framework

Chapter 3: Methodology

3.1 System Design

The final credit risk prediction system is architected as a complete, modular pipeline that reflects all stages of the machine learning lifecycle. This implementation is directly derived from an interactive Python notebook and structured to address the unique challenges of credit scoring in data-scarce environments. The design follows principles aligned with the **CRISP-DM** (Cross-Industry Standard Process for Data Mining) framework and supports full automation, fairness, and transparency.

The pipeline begins with the acquisition of two separate datasets: `Demographic Data` and `Credit Score Data`. These datasets are merged and undergo an extensive preprocessing phase, including handling of missing values, standardization, and one-hot encoding of categorical variables using the `pandas.get_dummies()` function.

A major design feature of the system is the use of **Conditional Tabular GAN (CTGAN)** for synthetic data generation. Financial attributes such as `INCOME`, `SAVINGS`, and `DEBT` are used to generate statistically realistic borrower profiles. Multiple dataset variants are created:

I. **Real-only dataset**: 1,000 genuine borrower records.

II. **Synthetic-only dataset**: 9,000 CTGAN-generated financial records.

III. **Mixed datasets**: Hybrids with 90%, 75%, and 50% real-to-synthetic ratios.

Feature engineering includes calculation of two key ratios: `Debt-to-Income` and `Debt-to-Savings`, both log-transformed via `np.log1p()` to reduce skewness. These features are added to enhance model learning and align with financial logic.

The modeling component incorporates four classifiers: `Logistic Regression`, `Random Forest`, `Gradient Boosting`, and `Neural Network (MLPClassifier)`. Models are evaluated using metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC, with results stored and compared across dataset variants.

To ensure fairness, sensitive demographic features including `sex`, `marital.status`, and `relationship` are excluded from training. Comparative results with and without these

features are analyzed. Explainability is provided through **SHAP (SHapley Additive Explanations)**, which identifies the most impactful features driving the model's predictions, confirming the reliance on financially relevant rather than demographic variables.

The final system includes an integrated deployment interface developed in **Streamlit**, where users can input borrower data and receive risk classifications in real-time. A downloadable PDF report is also generated using `FPDF`, allowing the result to be used operationally in field settings.

This system design effectively balances predictive accuracy, interpretability, and fairness, while also accommodating operational deployment in low-resource environments typical of many emerging markets.

The data science methodology adopted in this study is illustrated in Figure 3.1, based on the CRISP-DM framework by Chapman et al. (2000).

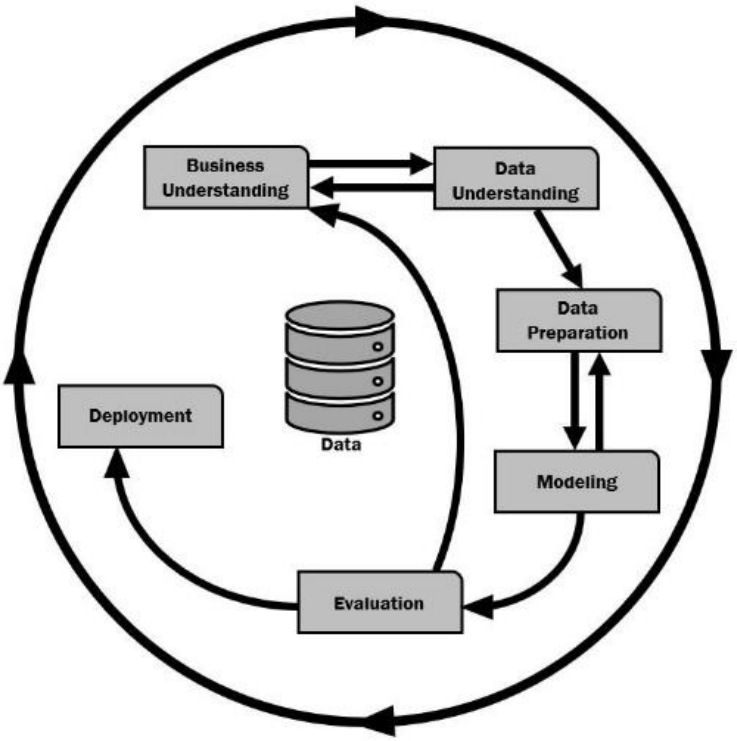


Figure 3.1: CRISP-DM Model for Data Mining

3.2 Business Understanding

Access to credit remains a critical enabler of economic mobility, yet many individuals in emerging markets are excluded due to the absence of formal financial histories. Tradi-

tional credit scoring systems, which rely heavily on structured financial records such as loan repayments or credit card usage, are ill-suited for regions where these records are sparse or entirely unavailable. This data scarcity limits financial institutions' ability to accurately assess creditworthiness, leading to a significant population being classified as "unscorable."

The primary business objective of this study is to develop a fair, scalable, and accurate credit scoring system that remains effective even in the absence of comprehensive financial histories. To achieve this, the research leverages a dual-source dataset architecture:

I. Demographic attributes (e.g., education, occupation, relationship role) which are more readily available across underbanked populations.

II. Synthetic financial profiles generated via Conditional Tabular GAN (CTGAN) to augment limited real-world data, simulating broader behavioral trends in borrower profiles.

By integrating synthetic data into the modeling pipeline, the system aims to mitigate data scarcity while maintaining predictive integrity and fairness. The design ensures that sensitive features such as sex and marital status are excluded from the training dataset, aligning with ethical AI principles and regulatory expectations around bias mitigation.

This study directly supports the strategic goal of expanding financial inclusion, particularly for populations historically marginalized by traditional credit systems. Furthermore, it positions credit-granting institutions to adopt machine learning approaches that are both practical and socially responsible, enabling smarter lending without perpetuating systemic bias.

3.3 Data Sources and Description

This study utilizes two primary structured datasets to construct and evaluate the proposed credit scoring models: a demographic dataset and a credit dataset. These datasets were curated and prepared to emulate the information typically accessible in resource-constrained financial environments, where complete credit histories are rarely available.

The initial demographic and financial datasets used in this study were obtained from Kaggle, a public data platform, and are both licensed under the CC0 Public Domain

license. These datasets were merged and curated to simulate real-world borrower information typical of microfinance and emerging market contexts. No personally identifiable information was included. All preprocessing and augmentation steps were performed on this composite dataset to ensure suitability for machine learning and fairness evaluation.

The **Demographic Dataset** comprises 1,000 anonymized individual records containing socio-economic attributes. Features include:

I. Education level

II. Marital status

III. Occupation category

IV. Relationship role within the household

V. Sex

These features are commonly captured during census-style or institutional surveys and were used during model development to simulate real-world data conditions.

The **Credit Dataset** contains financial attributes relevant for risk modeling, including:

I. Reported income

II. Total savings

III. Outstanding debt

IV. A binary indicator of credit default (1 = default, 0 = no default)

A random sample of 1,000 rows from the demographic dataset was aligned and merged with the credit dataset, creating a unified dataset of 1,000 entries. This merged dataset formed the basis for early model testing, feature engineering, and synthetic augmentation. For the main experimental phase, a CTGAN-based 10,000-row dataset was generated and used for the final modeling tasks. Table 3.1 summarizes the main groups of features used in the final 10K dataset.

Table 3.1: Summary of Feature Groups in the Main Modeling Dataset (10K)

Feature Group	**Sample Variables**	**No. of Variables**
Demographic Attributes	education, marital status, occupation, relationship, sex	5
Financial Indicators	income, savings, debt, default status	4
Engineered Indicators	R_DEBT_INCOME, R_DEBT_SAVINGS, CAT_DEBT, CAT_SAVINGS_ACCOUNT	4
Total	—	**13**

All features listed above were included during early and main modeling stages. To assess algorithmic bias and enhance fairness, sensitive demographic variables `sex` and `marital.status` were excluded in fairness-specific experiments. However, the final deployed model retains the `relationship` feature, as it demonstrated consistent predictive relevance without significant fairness trade-offs.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the data structure and inform preprocessing and modeling steps. The process is divided into the following stages:

3.4.1 Initial Data Inspection

The merged dataset was created by combining 1,000 sampled demographic records with financial records using `pandas.concat()`. To ensure data integrity:

I. Missing values were assessed using `merged_df.isnull().sum()`.

II. A heatmap was plotted using `sns.heatmap()` to visually confirm the absence of missing values.

3.4.2 Target Variable Distribution

To evaluate class balance, a count plot was created using `sns.countplot()` for the binary target variable `DEFAULT`. This plot helped assess whether default and non-default classes were imbalanced, which is crucial for training predictive models.

3.4.3 Distribution Analysis of Financial Variables

The distributions of key financial features—`INCOME`, `SAVINGS`, and `DEBT`—were visualized using histograms overlaid with KDE plots via `sns.histplot()`. These plots helped identify skewness and outliers, guiding the need for transformations such as logarithmic scaling.

3.4.4 Boxplot Analysis by Target Category

Boxplots for each financial variable were generated using `sns.boxplot()`, stratified by `DEFAULT` status. This allowed visual comparison of medians, variability, and outlier prevalence between defaulters and non-defaulters.

3.4.5 Correlation Matrix for Key Financial Indicators

To examine relationships among numerical variables, a correlation matrix was computed using `.corr()` for `INCOME`, `SAVINGS`, `DEBT`, `R_DEBT_INCOME`, and `R_DEBT_SAVINGS`. A heatmap rendered using `sns.heatmap()` enabled visual assessment of potential multicollinearity and associations between features.

3.5 Data Preprocessing and Transformation

3.5.1 Feature Engineering and Log Transformations

To enhance model performance and embed financial behavioral insights, two ratio-based features were engineered and included in the dataset:

I. **Debt-to-Income Ratio (`R_DEBT_INCOME`):** This was calculated as `DEBT / (INCOME + 1)` to quantify the proportion of income allocated to debt obligations.

II. **Debt-to-Savings Ratio (`R_DEBT_SAVINGS`):** This metric, computed as `DEBT / (SAVINGS + 1)`, reflects the extent to which savings can buffer against debt exposure.

To mitigate the effects of outliers and reduce feature skewness, both ratios were log-transformed using the `np.log1p()` function. This transformation ensured that the resulting values were scale-consistent, stable for zero or near-zero inputs, and better suited for machine learning models.

These two engineered features were retained in all subsequent modeling stages, including the final deployment pipeline.

3.5.2 One-Hot Encoding of Categorical Variables

To ensure compatibility with machine learning algorithms, all categorical features were transformed using one-hot encoding. This process converted each categorical variable into a set of binary indicators, with each column representing a distinct category.

Encoding was performed using the `pandas.get_dummies()` function with the parameter `drop_first=True`. This ensured that one category per feature was dropped to prevent multicollinearity while preserving complete feature representation.

The categorical variables subjected to this transformation included `education`, `marital.status`, `occupation`, `relationship`, and `sex`.

All resulting binary features were integrated into the training and testing datasets following imputation and formed part of the model input during the main development phase.

3.5.3 Numerical Standardization

To ensure that all numerical features were on comparable scales and contributed proportionately to model learning, standardization was applied across all continuous variables. This was especially critical given the varied ranges and units of measurement among financial indicators.

Standardization was implemented using the `StandardScaler` function , which transforms features to have zero mean and unit variance. The scaler was fitted on the training data and subsequently applied to both the training and test sets to prevent data leakage.

The standardized features included:

I. `INCOME`

II. `SAVINGS`

III. `DEBT`

IV. `R_DEBT_INCOME`

V. `R_DEBT_SAVINGS`

VI. `T_EXPENDITURE_12`

This process ensured feature scale consistency and improved the convergence and interpretability of scale-sensitive models such as logistic regression and multilayer perceptrons.

3.5.4 Train-Test Split Strategy

The dataset was split into training and testing subsets using a 70/30 ratio. Stratification was applied on the target variable (`DEFAULT`) to ensure proportional representation of both classes across the splits.

This procedure was implemented using the `train_test_split` function with a fixed random state for reproducibility. The split was conducted after all preprocessing steps, including categorical encoding and numerical standardization, forming the final datasets for model training and evaluation.

3.6 Synthetic Data Generation using CTGAN

One of the major barriers to effective credit scoring in emerging markets is the lack of sufficient structured financial data. In many cases, large populations operate outside formal financial systems, making it difficult to collect consistent financial records such as income, savings, and debt history. To address this challenge, this study employed a data augmentation strategy using **Generative Adversarial Networks (GANs)**, specifically the **Conditional Tabular GAN (CTGAN)**, to generate realistic synthetic financial data.

3.6.1 Overview of Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (Goodfellow et al., 2014b), were employed as the foundational technique for synthetic data generation in this study. GANs consist of two neural networks trained in opposition: a **Generator**, which creates synthetic data samples, and a **Discriminator**, which attempts to distinguish between real and generated data.

These networks are trained concurrently in a zero-sum game framework, where the generator aims to produce increasingly realistic samples that can deceive the discriminator,

while the discriminator seeks to correctly classify inputs as real or synthetic. This adversarial training strategy enables GANs to learn complex data distributions effectively.

Due to their high capacity for distribution learning, GANs are particularly suited for generating synthetic datasets in contexts with limited real data availability. In this study, a variant tailored for tabular data **CTGAN** was implemented to handle the structured, heterogeneous nature of financial and demographic data.

3.6.2 CTGAN Architecture and Justification

This study employed the **Conditional Tabular GAN (CTGAN)** for synthetic data generation, selected for its architectural suitability in handling the complexities of tabular datasets containing both continuous and categorical features. Unlike conventional GANs, CTGAN incorporates two key innovations:

I. A **conditional generator**, which generates samples conditioned on a discrete feature, improving control and balancing of categorical distributions.

II. A **mode-specific normalization technique**, which ensures consistent representation of continuous features across varying distributions.

The CTGAN architecture includes a generator that learns to produce realistic synthetic rows based on sampled category conditions, and a discriminator that assesses authenticity by jointly evaluating the sample and the conditioning information. Furthermore, CTGAN employs a *training-by-sampling* approach, giving underrepresented categories more sampling priority during training, thereby mitigating class imbalance.

This design aligns well with the project's dataset, which comprises a combination of sociodemographic categorical variables and financial numerical indicators. CTGAN's ability to handle such mixed data types effectively justified its selection for generating synthetic borrower profiles.

3.6.3 Initial Synthetic Data Generation

Prior to hyperparameter tuning, an initial synthetic dataset was generated using a baseline CTGAN configuration. This step validated the full augmentation pipeline and served as the first synthetic integration into the modeling process.

The CTGAN model was configured with the following parameters:

I. **Epochs:** 500

II. **Batch size:** 512

III. **PAC setting:** 1 (to ensure stable training on small datasets)

Training was conducted on a subset of 1,000 real financial records containing six variables: `INCOME`, `SAVINGS`, `DEBT`, `DEFAULT`, `CAT_DEBT`, and `CAT_SAVINGS_ACCOUNT`. The discrete variables were specified explicitly during training.

After model convergence, 10,000 synthetic financial records were generated, and a random sample of 9,000 was selected. Negative values in continuous fields were clipped for validity. Separately, 9,000 demographic entries were sampled from the original `demo_df` dataset and merged with the synthetic financial data to form fully synthetic borrower profiles.

The 9,000 synthetic profiles were concatenated with 1,000 real borrower records to construct a 10,000-row dataset. Feature engineering transformations were re-applied on the combined dataset to complete preprocessing before modeling.

3.6.4 Hyperparameter Tuning of CTGAN

To enhance the realism of generated data, a structured hyperparameter tuning process was carried out on the CTGAN model. The goal was to minimize statistical divergence between real and synthetic financial distributions using multiple training configurations.

The following parameter combinations were evaluated:

I. **Epochs:** 300, 500, 1000

II. **Batch sizes:** 256, 512

III. **Embedding dimensions:** 128, 256

All experiments used a `pac` setting of 1 to stabilize training on the limited dataset. Each configuration was trained on the 1,000-row financial dataset, and 1,000 synthetic samples were generated post-training.

To quantify similarity, the Kolmogorov–Smirnov (KS) test was applied to three continuous variables `INCOME`, `SAVINGS`, and `DEBT` comparing distributions between real and gen-

erated samples. The average KS-statistic across these features served as the performance metric.

The optimal configuration was found to be:

I. **Epochs:** 500

II. **Batch size:** 256

III. **Embedding dimension:** 256

This configuration produced the lowest average KS-statistic and was selected for the final training and generation of synthetic records.

3.6.5 Training the Final CTGAN Model

Following the hyperparameter tuning phase, the final CTGAN model was trained using the best-performing configuration.

The model was trained on the 1,000-row financial dataset, with discrete columns specified as `DEFAULT`, `CAT_DEBT`, and `CAT_SAVINGS_ACCOUNT`. Upon completion, the generator produced 10,000 synthetic financial records.

A subset of 9,000 records was selected, clipped for non-negative values, and index-reset. In parallel, 9,000 demographic rows were sampled from the original dataset and merged with the synthetic financial attributes to construct full synthetic borrower profiles.

These 9,000 synthetic profiles were combined with the original 1,000 real records to yield a final 10,000-row dataset. Feature engineering transformations were re-applied to this unified dataset to prepare it for modeling and deployment.

3.6.6 Synthetic Dataset Variants

To evaluate model performance under different data conditions, three dataset configurations were constructed:

I. **Real-Only Dataset:** Consisting of 1,000 real borrower records that included both financial and demographic features. This dataset served as a control baseline for comparison.

II. **Synthetic-Only Dataset:** Comprised of 9,000 synthetic borrower profiles generated by combining CTGAN-produced financial attributes with independently sampled demographic data.

III. **Hybrid Dataset (90% Synthetic + 10% Real):** Formed by merging 9,000 synthetic profiles with the 1,000 real records, resulting in a 10,000-row hybrid dataset used for final training and model deployment.

All datasets maintained a unified schema, ensuring compatibility in preprocessing, feature engineering, and downstream evaluation workflows. This structure enabled rigorous testing across multiple data availability scenarios.

3.6.7 Validation of Synthetic Data Quality

The statistical fidelity of the generated synthetic data was evaluated using the Kolmogorov–Smirnov (KS) test, which measures the divergence between the empirical distributions of real and synthetic datasets.

This analysis focused on three critical financial indicators:

I. `INCOME`

II. `SAVINGS`

III. `DEBT`

KS-statistics were computed during the CTGAN tuning process to compare each synthetic sample set against the original real dataset. The test provided a robust quantitative measure of similarity in distribution shape and spread.

The final CTGAN model was selected based on its ability to produce the lowest average KS-statistic across the three variables. This approach ensured that the synthetic financial data used for training and evaluation closely resembled the statistical characteristics of real-world borrower data.

3.7 Model Development and Evaluation

This section outlines the machine learning models selected for this study, as well as the evaluation metrics employed to measure their performance in predicting credit default.

All models were trained on the 10,000-row hybrid dataset under consistent preprocessing and feature engineering conditions.

3.7.1 Machine Learning Models

Four supervised machine learning models were employed in this study. These were chosen based on their widespread use in credit risk modeling, their balance between interpretability and complexity, and their suitability for structured tabular data:

I. **Logistic Regression**

Logistic Regression is a linear model that estimates the probability of default using a logistic function:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

It was included as a baseline model due to its simplicity and ease of interpretation.

II. **Random Forest**

Random Forest is an ensemble model that aggregates the output of multiple decision trees via majority voting:

$$\hat{y} = \text{mode}(T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_k(\mathbf{x}))$$

It is robust to overfitting and well-suited for both linear and non-linear relationships.

III. **Gradient Boosting**

Gradient Boosting builds trees sequentially, each one correcting the residuals of its predecessor:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma h_m(\mathbf{x})$$

where γ is the learning rate and h_m is the new weak learner. It is highly effective on structured data.

IV. **Neural Network**

A feedforward neural network was used to capture complex nonlinear patterns in the data. It comprised an input layer, multiple hidden layers with ReLU activations, and a sigmoid-activated output layer:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Weight optimization was performed via backpropagation using binary cross-entropy loss.

3.7.2 Evaluation Metrics & Formulas

To ensure robust evaluation of each classifier, multiple performance metrics were used. These metrics are particularly important in credit scoring, where class imbalance is common and misclassifications carry different costs.

I. **Accuracy** — Proportion of total correct predictions:

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

II. **Precision** — Proportion of predicted defaults that are actual defaults:

$$\text{Precision} = \frac{TP}{TP + FP}$$

III. **Recall (Sensitivity)** — Proportion of actual defaults correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN}$$

IV. **F1 Score** — Harmonic mean of precision and recall:

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

V. **AUC-ROC** — Area under the Receiver Operating Characteristic curve, which

reflects the model's ability to distinguish between classes across all thresholds.

3.7.3 Baseline Model Setup

To establish a baseline for model performance, all selected machine learning algorithms were first trained on the original 1,000-row dataset containing only real borrower records. This baseline setup reflects real-world scenarios where limited financial data is available, especially in emerging markets.

A stratified 70:30 train-test split was employed to maintain the class distribution of default and non-default labels across both subsets. Preprocessing procedures—including one-hot encoding, standardization, and log-transformed feature engineering—were applied identically to both training and test data.

No synthetic data, oversampling techniques, or class reweighting strategies were introduced at this stage. These baseline runs were used to compare against later models trained on synthetic and hybrid datasets.

3.7.4 Training on Synthetic Data (Full 10K)

To assess model performance in a purely synthetic setting, all four selected machine learning models—Logistic Regression, Random Forest, Gradient Boosting, and Neural Network—were trained using a fully synthetic dataset containing 10,000 records.

This dataset was created by generating 9,000 synthetic financial records using the final CTGAN model and combining them with 9,000 randomly sampled demographic records. These were then appended to the original 1,000 real merged records, producing a full dataset of 10,000 rows.

The same preprocessing pipeline used for the real and mixed datasets was applied to this synthetic dataset. This included:

I. Feature engineering to compute debt-to-income and debt-to-savings ratios

II. Log transformations to normalize skewed features

III. One-hot encoding of categorical variables

IV. Standardization of numerical features

The dataset was then split into training and test sets using a 70:30 stratified ratio, ensuring the class distribution of defaulters and non-defaulters remained consistent across both subsets.

Each model was independently trained on this fully synthetic dataset to simulate performance in environments where real financial data is entirely unavailable.

3.7.5 Training on Mixed Datasets (90/10, 75/25, 50/50)

To evaluate the impact of combining real and synthetic data on model performance, three mixed datasets were created with varying proportions of synthetic to real records. These hybrid datasets consisted of:

I. 90% real data and 10% synthetic data

II. 75% real data and 25% synthetic data

III. 50% real data and 50% synthetic data

Synthetic financial data was generated using the final tuned CTGAN model. These synthetic financial records were merged with randomly sampled demographic entries from the original dataset to simulate realistic borrower profiles. Each merged dataset was constructed to contain a total of 1,000 records.

The following preprocessing steps were applied uniformly across all mixed datasets:

I. Feature engineering (debt-to-income and debt-to-savings ratios)

II. Log transformation of ratio features

III. One-hot encoding of categorical variables

IV. Standardization of numerical features

V. Missing value imputation (numeric and categorical)

Each dataset was then split into training and test sets using a 70:30 stratified strategy to maintain class balance. All four machine learning models were independently trained on each dataset configuration using the same preprocessing and evaluation procedures.

This approach facilitated a robust comparison of model performance across varying levels of synthetic augmentation.

3.7.6 Final Model Performance Summary

To enable a comparative analysis of all modeling configurations, evaluation metrics from each training scenario were consolidated into a unified summary. This included models trained on the following datasets:

I. Real-only dataset (1,000 rows)

II. Fully synthetic dataset (10,000 synthetic + real hybrid rows)

III. Mixed datasets:

   a. 90% real / 10% synthetic

   b. 75% real / 25% synthetic

   c. 50% real / 50% synthetic

Each dataset underwent identical preprocessing procedures, including feature engineering, log transformation, one-hot encoding, standardization, and stratified splitting.

After training, core performance metrics—Accuracy, Precision, Recall, F1-Score, and AUC-ROC—were computed for all models under each data configuration. These metrics were then aggregated into a comparative table to enable structured evaluation of each modeling approach.

This consolidation was instrumental in identifying the best-performing model configuration for subsequent tuning and deployment.

3.8 Final Model Optimization

This section outlines the optimization and interpretability strategies applied to the best-performing model, which was a Random Forest trained on the 75/25 mixed dataset. The process aimed to refine model performance, evaluate fairness, and enhance transparency through explainability tools.

3.8.1 Hyperparameter Tuning of Selected Model (Random Forest)

To improve model generalization and predictive accuracy, a hyperparameter tuning procedure was performed on the Random Forest classifier using the training portion of the

75/25 mixed dataset. This dataset was selected based on prior performance comparisons across various data configurations.

`GridSearchCV`, a comprehensive grid-based search approach combined with `5-fold cross-validation`, was employed to identify the optimal set of hyperparameters. The model was evaluated using the **F1-score**, which balances precision and recall and is particularly suited for imbalanced classification problems such as credit default prediction.

The following hyperparameters were tuned:

I. `n_estimators` — Number of decision trees in the ensemble

II. `max_depth` — Maximum depth of individual trees

III. `min_samples_split` — Minimum number of samples required to split an internal node

IV. `min_samples_leaf` — Minimum number of samples required to be at a leaf node

All models were configured with a fixed random seed for reproducibility. Parallelization was achieved through `n_jobs=-1` to optimize computational efficiency.

The best-performing hyperparameter configuration identified through GridSearchCV was subsequently used to retrain the Random Forest model on the full 75/25 mixed training set.

3.8.2 Retraining on Full Dataset

Following hyperparameter tuning, the optimal configuration was applied to retrain the Random Forest model using the entire training set derived from the 75/25 mixed dataset. This retraining phase was designed to fully leverage the combined real and synthetic records, ensuring maximum learning capacity from the available data.

To address the class imbalance inherent in credit default datasets, the model was trained with the `class_weight='balanced'` parameter. This adjustment reweights the classes inversely proportional to their frequency, preventing the model from being biased towards the majority class.

The retrained model served as the finalized candidate for downstream evaluations including statistical confidence assessments, fairness checks, and explainability analysis.

3.8.3 Bootstrapped Confidence Intervals

To assess the statistical reliability and stability of the final Random Forest model's performance, bootstrapping was applied to compute 95% confidence intervals for key evaluation metrics. Bootstrapping is a resampling technique that allows estimation of the variability of metrics by repeatedly sampling, with replacement, from the prediction results.

For this analysis, 1,000 bootstrap resamples were drawn from the test data. The following performance metrics were evaluated across these samples:

I. F1-Score

II. Recall

III. AUC-ROC

For each metric, the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution were reported as the bounds of the 95% confidence interval. This statistical insight strengthens the robustness claims of the model and confirms the reliability of its predictive capabilities across different samples.

3.8.4 Fairness Analysis – Exclusion of Sensitive Variables

To evaluate the ethical integrity of the credit scoring system, a fairness analysis was conducted by removing sensitive demographic attributes that could potentially introduce bias. Specifically, the variables `sex`, `marital.status`, and `relationship` were excluded from the feature set used for training and testing.

The same data preprocessing pipeline was applied to this reduced dataset to maintain consistency in transformation, encoding, and scaling. A new Random Forest model was trained using this fair dataset configuration, with the goal of assessing whether performance could be preserved without reliance on potentially discriminatory features.

This fairness-aware model was then evaluated using the same metrics and methodology as the full model. The comparison provided insights into the trade-off between ethical

compliance and predictive power, supporting the design of an inclusive and responsible AI system for credit scoring.

3.8.5 SHAP-based Explainability

To enhance transparency and interpretability, SHAP (SHapley Additive exPlanations) was employed to analyze the predictions made by the final fairness-aware Random Forest model. SHAP values are grounded in cooperative game theory and quantify each feature's contribution to the model's output for a given prediction.

Global interpretability was achieved using SHAP summary plots, which rank features by their average impact on the model's predictions. These plots indicate both the magnitude and direction of influence, helping stakeholders understand which features consistently drive the model towards a classification of default or non-default.

SHAP also provides local interpretability by explaining individual predictions, making it suitable for case-level analysis in credit scoring applications. This is particularly important in scenarios involving high-stakes decisions, such as loan approvals, where transparency is critical for regulatory and ethical considerations.

This approach was intended to verify that the model prioritizes logically relevant variables such as financial indicators over demographic factors, supporting ethical and explainable AI practices.

3.9 Model Deployment

To ensure real-world applicability, the final credit scoring model was integrated into an interactive web-based application using Streamlit, a Python framework designed for rapid prototyping of machine learning tools. This deployment phase marked the transition from research to operational readiness, allowing the model to be accessed and used by non-technical stakeholders such as credit officers.

3.9.1 Deployment Objective

The main objective was to transform the trained machine learning pipeline into a responsive decision-support system. By offering a user-friendly interface, the application allows financial practitioners to input borrower characteristics and receive risk assess-

ments instantly. This deployment not only extends the model's reach but also enhances transparency and adoption by embedding predictive insights into familiar workflows.

3.9.2 Tools and Deployment Stack

The deployment environment was designed to ensure smooth integration, model reproducibility, and real-time prediction capabilities. The following tools and technologies formed the core of the deployment stack:

I. **Python 3.10+** – Provided the foundational programming environment for the model and app logic.

II. **Scikit-learn** – Used for training the Random Forest classifier and generating predictions.

III. **Joblib** – Enabled serialization and loading of the trained model and associated feature set.

IV. **Pandas and NumPy** – Powered data manipulation and feature engineering processes within the app.

V. **Streamlit** – Served as the primary framework for building the user interface and hosting the application.

VI. **Streamlit Cloud** – Provided a cloud-based hosting solution, ensuring that the tool could be accessed remotely by end users without local dependencies.

3.9.3 Streamlit Application Structure

The Streamlit application was designed with an emphasis on usability and clarity, enabling seamless interaction for non-technical users. The interface is organized into four navigable pages:

I. **Home Page** – Provides a high-level overview of the app, its purpose, and key disclaimers.

II. **Prediction Page** – Allows users to input borrower data, trigger the model, and receive credit risk predictions.

III. **Model Explanation Page** – Offers transparency into the logic behind predictions, highlighting fairness considerations and model limitations.

IV. **Disclaimer Page** – Reinforces that the tool is intended for educational and demonstration purposes only.

The app embeds all preprocessing logic (e.g., log transformations, categorical encoding, and feature scaling) directly within the prediction pipeline. This ensures that the user inputs are processed consistently with the model's original training setup, maintaining alignment with expected data formats and logic.

3.9.4 Model Integration and Prediction Logic

The deployed application integrates the final fairness-aware Random Forest classifier using serialized components. Both the model file and its corresponding feature list were saved using `joblib` and loaded dynamically upon app initialization. This setup ensures that model inference is performed in real time as users input borrower data.

Once user inputs are received via the interface, the following steps are executed:

I. Financial ratios (e.g., debt-to-income and debt-to-savings) are computed and log-transformed.

II. Binary flags for debt and savings account presence are generated.

III. Categorical features such as education, occupation, and household relationship role are encoded using one-hot encoding aligned with the model's training features.

IV. All transformed features are mapped to the model's expected schema, and any missing features are zero-filled.

The processed input is then passed to the model for prediction. The predicted probability is compared against a user-defined threshold (default: 0.40) to classify the borrower into "High Risk" or "Low Risk" categories. Additionally, a PDF summary report of the prediction is auto-generated, allowing for recordkeeping and offline sharing.

3.9.5 User Workflow and Experience

The application was designed to provide a seamless and intuitive user experience. The complete interaction workflow is structured as follows:

I. The user accesses the deployed application via a secure link hosted on Streamlit Cloud.

II. Through an interactive form, the user inputs borrower-specific details including income, savings, debt, education level, occupation, and household relationship status.

III. Upon clicking the `Predict` button, the system performs real-time inference using the preloaded Random Forest model.

IV. The prediction result whether the borrower is at high or low risk of default is displayed alongside the associated probability and a contextual risk band (e.g., "Medium Risk").

V. The user is given the option to download a PDF report summarizing the prediction, borrower input features, and explanation notes.

This real-time feedback loop enhances operational decision-making and aligns predictive analytics with practical credit risk assessment needs. The interface is designed to be lightweight, secure, and accessible even in low-resource settings.

Chapter 4: System Design and Architecture

4.1 Introduction

This chapter presents the full design and architecture of the developed credit scoring system. It details both the offline machine learning pipeline and the interactive web-based deployment. The system was structured to provide real-time credit risk predictions by integrating robust data science practices with user-friendly accessibility.

The architecture covers the entire data science lifecycle, starting from data acquisition, preprocessing, model training, evaluation, and culminating in model deployment. Once the optimal model was identified, it was deployed through a Streamlit application, allowing end-users to interact with the model through a simple web browser.

The chapter is organized to describe the system's major components, data flow, technology stack, and user interaction process. It also highlights the integration between offline development and online prediction environments, ensuring a seamless and scalable solution for credit risk assessment.

4.2 System Overview

The credit scoring system was designed as a modular, end-to-end pipeline that seamlessly integrates offline model development with online prediction deployment. It comprises two main environments: an offline environment for model training and testing, and an online environment for real-time risk scoring via a web interface.

The architecture is composed of the following layered components:

I. **Data Layer** — Stores the original demographic dataset and the synthetic financial data used for training and experimentation.

II. **Preprocessing Layer** — Implements data cleaning, feature engineering, encoding, and scaling. All transformations are mirrored during deployment to ensure consistency.

III. **Model Layer** — Hosts various machine learning models including Logistic Regression, Random Forest, Gradient Boosting, and Neural Network. The best model, Random Forest, was selected for deployment.

IV. **Prediction Layer** — Loads the serialized model and applies it to new inputs to generate real-time credit risk predictions.

V. **User Interface Layer** — Built using Streamlit, this interface enables user interaction through an intuitive web form.

VI. **Deployment Environment** — Hosted on Streamlit Cloud, making the app accessible remotely via browser.

The technologies used throughout the system's development, training, and deployment pipeline are outlined in Table 4.1.

Table 4.1: System Technology Stack

Component	**Technology/Tool Used**
Data Processing	Python (Pandas, NumPy)
Model Training	Scikit-learn
Model Serialization	Joblib
Deployment	Streamlit, Streamlit Cloud
Platform	Google Colab

This layered design ensures flexibility, modularity, and scalability, making the system suitable for integration into microfinance workflows where data scarcity and operational simplicity are essential.The complete architecture of the system is illustrated in Figure 4.1.

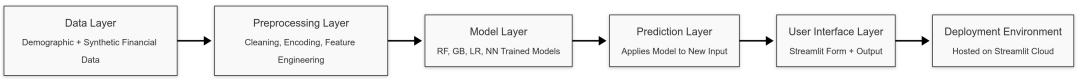


Figure 4.1: System Architecture Based on Layered Pipeline Design

4.3 Offline Pipeline Design

The offline pipeline encompasses the full machine learning workflow executed prior to deployment. It focuses on dataset construction, preprocessing, model training, evaluation, and serialization. This environment was implemented entirely in Python via Google Colab, enabling reproducibility and ease of experimentation.

The stages of the offline pipeline include:

I. **Data Acquisition**

A merged dataset was created by combining 1,000 real demographic records with corresponding financial variables. To address data scarcity, 9,000 synthetic financial records were generated using CTGAN and merged with sampled demographic data to create a comprehensive 10,000-row dataset.

II. **Data Preprocessing**

The preprocessing steps applied include:

a. **Missing Value Handling** – Imputation using appropriate statistical methods.

b. **Feature Transformation** – Logarithmic transformations (e.g., `log1p`) were applied to skewed ratio variables such as Debt-to-Income and Debt-to-Savings.

c. **Categorical Encoding** – All categorical variables were one-hot encoded.

d. **Standardization** – Numerical features were standardized using `StandardScaler` to ensure zero mean and unit variance.

III. **Feature Engineering**

Domain-specific features were derived to enhance model interpretability and predictive power:

a. **Financial Ratios** – Calculated Debt-to-Income and Debt-to-Savings ratios.

b. **Log-transformed Ratios** – To reduce skewness and the influence of outliers.

IV. **Synthetic Data Construction**

Synthetic financial variables were generated using CTGAN, which was trained on real financial data. Multiple synthetic dataset variants were created by mixing synthetic and real records at varying ratios (90/10, 75/25, 50/50).

V. **Model Training**

Four classifiers were trained on each dataset variant:

a. Logistic Regression

b. Random Forest

c. Gradient Boosting

d. Neural Network

VI. **Model Evaluation**

Each model was assessed using Accuracy, Precision, Recall, F1-Score, and AUC-ROC across real, synthetic, and hybrid datasets to identify the most effective combination.

VII. **Model Selection and Serialization**

The Random Forest model trained on the 75/25 mixed dataset emerged as the top performer. It was serialized using `joblib` for deployment in the Streamlit application.

4.4 Online System Architecture

The online architecture transforms the offline-trained model into a real-time credit scoring application. It is implemented using Streamlit and hosted on Streamlit Cloud, allowing remote access through a browser without requiring technical expertise or local installations.

The components of the online system are structured as follows:

I. **Streamlit Interface (Frontend)**

The user interacts with the system through an intuitive, browser-based form. Users can input borrower information such as income, savings, debt, education level, occupation, and household relationship role. The layout is simple and optimized for decision support.

II. **Input Processing Layer**

Upon form submission, input data is passed to the backend module where it undergoes transformations that mirror those used during training. This includes:

a. Feature Engineering (ratios)

b. Encoding of categorical variables

c. Standardization of numerical features

III. **Model Loading and Prediction Engine**

The deployed model a tuned and fairness aware Random Forest—is loaded using `joblib`. Incoming data is processed and passed to the model, which outputs a binary risk prediction (High or Low Risk).

IV. **Output Display Module**

The prediction is displayed immediately in the web interface, providing the user with clear feedback on the borrower's risk status. Additional confidence metrics such as risk probability and risk bands are also shown.

V. **Hosting Environment**

The full application is hosted on Streamlit Cloud. This enables accessibility, eliminates the need for local infrastructure, and allows for seamless interaction across multiple users and devices.

This setup bridges the gap between machine learning research and practical use, ensuring transparency, consistency, and accessibility in deploying a credit risk model for real-world applications.

4.5 Data Flow Design

The data flow architecture outlines how borrower inputs are transformed into credit risk predictions within the deployed application. The design ensures that the processing logic used during model development is faithfully reproduced during real-time prediction, maintaining consistency and accuracy.

The key stages in the system's data flow are as follows:

I. **User Input Collection**:

Users provide borrower information such as income, savings, debt, education, occupation, and relationship role through the web-based Streamlit form.

II. **Data Transmission**:

Once the user submits the form, the input data is sent to the backend where it undergoes processing. Streamlit handles this transmission securely and efficiently.

III. **Preprocessing and Feature Engineering**:

The backend replicates the transformations applied during the training phase, which

include:

a. Feature Engineering (e.g., calculating debt-to-income ratios)

b. One-hot encoding of categorical variables

c. Scaling of numerical features

IV. **Model Inference**:

The preprocessed input is fed into the loaded Random Forest model. The model produces a probability score indicating the likelihood of credit default.

V. **Output Presentation**:

Based on a configurable threshold (e.g., 0.4), the system classifies the borrower as either "High Risk" or "Low Risk." The prediction and associated risk band are displayed instantly on the user interface.

This real-time data flow ensures a seamless and consistent experience, tightly coupling the offline machine learning pipeline with the live online environment.

This data flow ensures a seamless integration between the offline modeling process and the online prediction environment. The sequence diagram in Figure 4.2 illustrates this interaction.

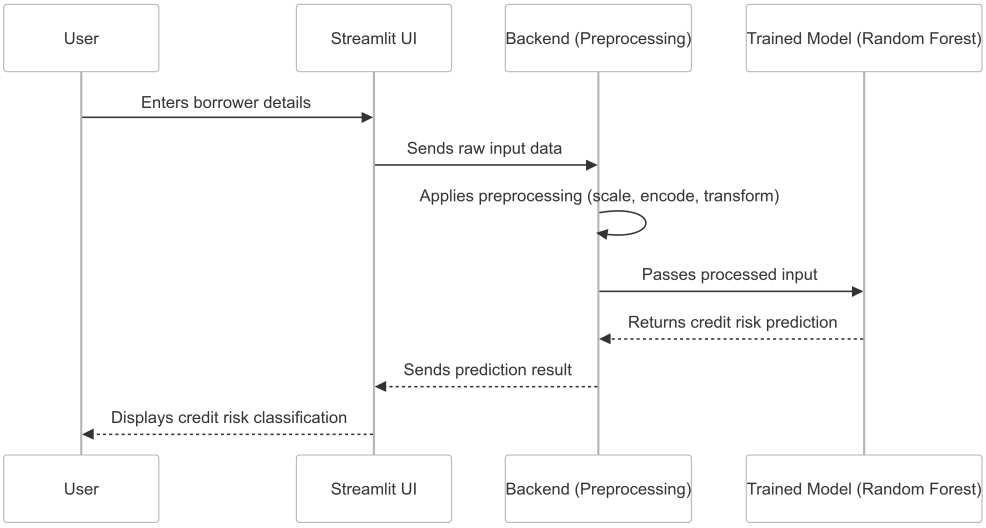


Figure 4.2: Sequence Diagram Illustrating Interaction Between the Offline-Trained Model and Online Prediction System

4.6 User Interface Design

The user interface (UI) of the deployed application was designed with a focus on simplicity, usability, and transparency. Built using Streamlit, it enables loan officers and financial analysts to easily interact with the credit scoring model via a browser-based form without requiring technical expertise.

The interface is logically organized into the following sections:

I. **Header Section**

Displays the system title (*"Credit Default Risk Assessment Tool"*) and a short description. It introduces the application's purpose and emphasizes responsible use.

II. **Input Form**

The form allows users to input borrower-specific information required to generate a credit risk prediction:

a. **Income**: Total reported monthly income (numeric input).

b. **Savings**: Total reported savings (numeric input).

c. **Debt**: Current outstanding debt (numeric input).

d. **Education Level**: Dropdown menu to select the borrower's highest educational attainment.

e. **Occupation**: Dropdown menu to select the borrower's primary occupation.

f. **Household Role**: Dropdown menu indicating the borrower's relationship within the household (e.g., Single, Married, Divorced).

g. **Threshold Slider**: Allows users to adjust the risk classification threshold.

Note: Sensitive features such as gender and marital status were deliberately excluded to uphold fairness principles.

III. **Prediction Trigger**

After completing the form, the user clicks the *"Predict Default Risk"* button to submit the input data. The backend processes the information, applies preprocessing steps, and uses the trained Random Forest model to generate a prediction.

IV. **Output Display**

The application presents the following results:

a. Predicted probability of default (expressed as a percentage).

b. Final risk classification ("High Risk" or "Low Risk").

c. Risk band categorization (Low, Medium, High).

d. Option to download a **PDF report** summarizing the borrower input and prediction results.

V. **Deployment and Accessibility**

The application is deployed on **Streamlit Cloud**, making it remotely accessible through a secure web URL. This cloud-based deployment ensures real-time usage, scalability, and device independence without requiring local installation.

Figure 4.3 illustrates the data flow through the user interface of the deployed credit scoring application, highlighting how input features are processed and passed to the model for real-time prediction.

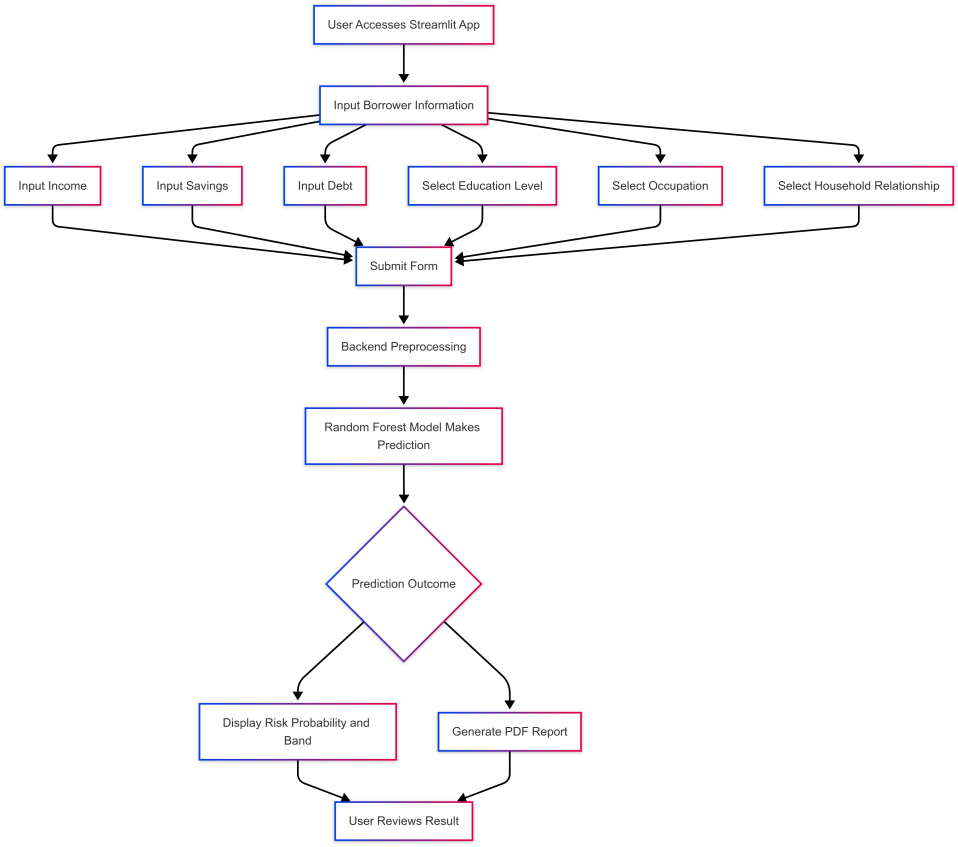


Figure 4.3: User Interface Data Flow in Credit Scoring Application

Chapter 5: System Implementation and Testing

5.1 Introduction

This chapter details the implementation and testing of the deployed credit scoring system developed in this study. The system integrates a trained machine learning model with a user-friendly, web-based interface to provide real-time borrower risk classification. The final solution is accessible via a browser, enabling financial analysts and credit officers to obtain risk predictions without technical expertise.

The implementation phase involved transforming the trained Random Forest model into an interactive and deployable tool using the Streamlit framework. This required setting up the prediction environment, ensuring consistent preprocessing steps were applied to user inputs, and integrating backend logic with a responsive frontend.

Furthermore, the chapter outlines the core functionalities built into the system, the testing methodologies used to ensure system reliability and usability, and an evaluation of how well the implemented solution meets the original project objectives. The aim was to deliver a practical and ethical decision-support tool for real-world credit risk assessment in data-constrained environments.

5.2 System Implementation

This section describes the complete setup and logic behind the implementation of the credit scoring system. The goal was to convert the best-performing machine learning model into a deployable, interactive application accessible through a browser.

5.2.1 Implementation Environment

The system was developed using accessible, open-source tools to support rapid prototyping and reproducibility. All machine learning workflows were executed in Google Colab, while deployment was handled via Streamlit Cloud. The tools and technologies used include:

I. **Programming Language**: Python 3.10

II. **Libraries and Tools**:

a. **Data Handling** – `Pandas`, `NumPy`

b. **Machine Learning** – `Scikit-learn`

c. **Model Persistence** – `Joblib`

d. **PDF Generation** – `FPDF`

e. **Frontend and Deployment** – `Streamlit`, hosted on Streamlit Cloud

5.2.2 Backend Implementation

The backend logic handles input preprocessing and prediction using the trained Random Forest model. Key responsibilities of the backend include:

I. Loading the serialized model and feature list from local storage using `joblib`

II. Performing feature engineering on the input, such as calculating:

a. Debt-to-Income Ratio

b. Debt-to-Savings Ratio

III. One-hot encoding of categorical variables (education, occupation, relationship)

IV. Ensuring all user inputs align with the feature schema used during training

V. Generating probability scores and final credit risk classifications

5.2.3 Frontend Implementation (Streamlit Interface)

The user interface was developed using Streamlit, designed for accessibility and responsiveness. It includes:

I. **Navigation Sidebar**: Allows users to toggle between different app pages:

a. Home

b. Predict

c. Model Explanation

d. Disclaimer

II. **Prediction Page**: Enables users to input financial and demographic data via:

a. Number fields for income, savings, and debt

b. Dropdowns for education, occupation, and relationship status

c. Risk threshold slider

III. **Prediction Output**: Displays classification, probability, and risk level

IV. **PDF Export Feature**: Generates a downloadable credit risk report

The modular architecture ensures separation of concerns and simplifies future upgrades, such as integrating new models or adding multilingual support.

5.3 Key Functionalities Implemented

This section outlines the core features that were successfully integrated into the deployed credit scoring system. These functionalities ensure that the model operates efficiently, delivers accurate predictions, and offers a seamless and intuitive user experience from data input to output generation.

5.3.1 Credit Risk Prediction Module

The primary function of the system is to classify borrowers into either a **"High Risk"** or **"Low Risk"** credit category using a trained Random Forest model. Once the user submits their details through the input form, the model immediately computes and displays the corresponding credit risk classification.

Key input features utilized by the model include:

I. Income

II. Savings

III. Debt

IV. Debt-to-Income Ratio (automatically calculated)

V. Debt-to-Savings Ratio (automatically calculated)

VI. Education level

VII. Occupation

VIII. Household relationship status

These features were selected based on their predictive relevance and consistency with the training dataset.

5.3.2 Real-Time Prediction Trigger

Predictions are generated in real time when the user clicks the **"Predict"** button. The backend system automatically applies all necessary preprocessing transformations to the input data, ensuring that the features align precisely with the model's expectations. This real-time capability makes the application highly suitable for environments requiring quick decision support, such as microfinance operations.

5.3.3 Frontend-Backend Integration

The system ensures a direct mapping between each user interface field and its corresponding backend model feature. The preprocessing pipeline used during model training including scaling, encoding, and feature engineering is faithfully replicated during the prediction phase. This strict pipeline consistency guarantees accurate, reliable model behavior.

5.3.4 Stateless Deployment

The system has been deployed in a fully stateless manner. No personally identifiable information (PII) is stored on the server after prediction. Each prediction session is isolated and independent, enhancing user privacy and aligning with ethical and data protection standards.

5.3.5 Lightweight and Cloud-Based Accessibility

The final application was deployed using **Streamlit Cloud**, making it accessible from any internet-connected device without requiring local installation. This lightweight deployment approach ensures scalability, broad accessibility, and practicality for deployment in low-resource environments.

5.4 User Interface

The user interface (UI) of the credit scoring system was designed with a strong focus on clarity, usability, and responsiveness. Built using Streamlit, it ensures that even non-technical users, such as loan officers and credit analysts, can seamlessly interact with the system and obtain credit risk predictions.

5.4.1 Header and Introduction Section

At the top of the application, users are greeted with the title **"Credit Default Prediction App"** and a short introductory description outlining the purpose of the tool. This immediately orients the user and clarifies the app's objective.

5.4.2 Input Form

The input form is the core component where users submit borrower information. It includes:

I. **Numerical Inputs**

   a. Income

   b. Savings

   c. Debt

II. **Categorical Inputs**

   a. Education Level (dropdown)

   b. Occupation (dropdown)

   c. Household Role (dropdown)

III. **Threshold Slider**

   a. A slider allows users to adjust the decision threshold for classifying credit risk, defaulting to 0.4.

All fields are mandatory to ensure complete data submission, and validation checks are in place to prevent submission of unrealistic values (e.g., all zeros for financial figures).

5.4.3 Prediction Trigger

Once the input form is completed, users click the **"Predict"** button. This triggers the backend to:

a. Apply feature engineering (calculate debt ratios)

b. Preprocess the input features

c. Generate a probability score

d. Classify the borrower into a "High Risk" or "Low Risk" category

5.4.4 Prediction Output Display

The prediction result is displayed instantly and clearly, with supportive visual cues:

a. **High Risk:** Red-colored alert

b. **Low Risk:** Green-colored success box

c. **Predicted Probability** and **Risk Band** (Low, Medium, High) are also displayed for better decision support.

Additionally, a downloadable PDF report is generated, summarizing the borrower's input information and the prediction results.

5.4.5 Navigation Menu

A sidebar navigation menu allows users to quickly move between the following pages:

a. Home: Introduction and overview

b. Predict: Credit risk prediction form

c. Model Explanation: Information about how the model works and its fairness considerations

d. Disclaimer: Usage warnings and contact details

5.4.6 Accessibility and Responsive Design

The app is fully accessible through any modern browser without installation. The layout automatically adjusts to different device sizes, ensuring usability across desktops, laptops, and mobile devices.Figures 5.1 and 5.2 display the two-step Streamlit input form, capturing borrower financial attributes and categorical fields. The resulting prediction interface is shown in Figure 5.3, where users are presented with a credit risk classification and corresponding probability.

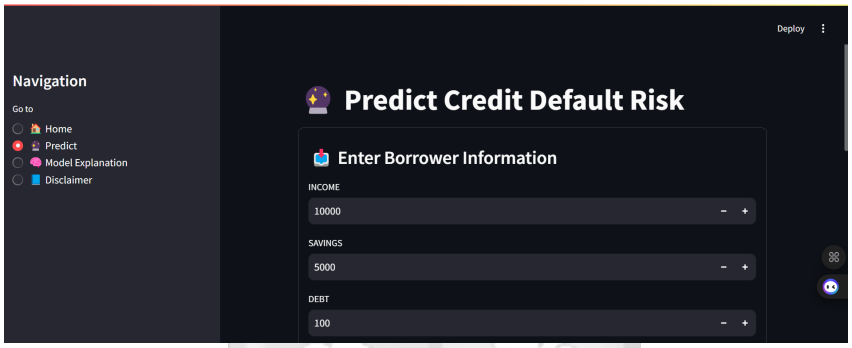


Figure 5.1: First part of the Streamlit input form showing borrower financial fields (Income, Savings, Debt).

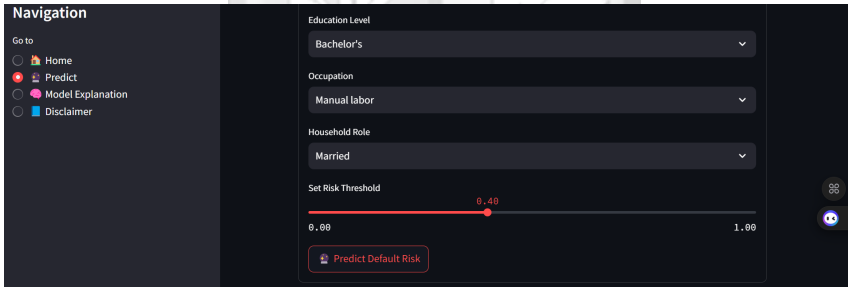


Figure 5.2: Second part of the Streamlit input form showing categorical selections and risk threshold adjustment.

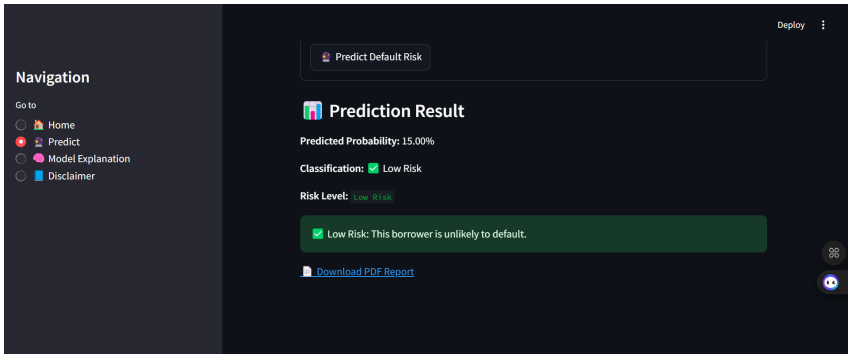


Figure 5.3: Prediction output displayed by the system indicating borrower credit risk category and probability.

5.5 System Testing

This section outlines the various testing approaches used to validate the performance, usability, and reliability of the deployed credit scoring system. These tests ensure that the system is not only technically functional but also accessible and intuitive for non-technical users such as credit officers in microfinance institutions.

5.5.1 Functional Testing

Functional testing was performed to verify that all core components of the system work as expected. The following aspects were tested:

I. The user input form correctly accepts numerical and categorical inputs.

II. The "Predict Credit Risk" button triggers the backend model pipeline.

III. The model processes inputs and returns either a "High Risk" or "Low Risk" prediction.

IV. Appropriate feedback is given for both valid and invalid inputs.

5.5.2 Usability Testing

Usability testing focused on ensuring the application is easy to navigate, even for users with limited technical knowledge. Informal peer reviews were conducted, where test users interacted with the application and provided feedback.

Key usability highlights:

I. Clear labeling of input fields.

II. Intuitive flow from input to output.

III. Use of color-coded outputs (green for low risk, red for high risk) for easier interpretation.

5.5.3 Compatibility Testing

To ensure broad accessibility, the application was tested across different platforms and devices:

I. Browsers tested: Google Chrome, Mozilla Firefox, and Microsoft Edge.

II. Devices tested: Desktop (Windows), Laptop, and Android mobile phone.

5.5.4 Navigation Testing

Navigation testing ensured smooth transitions and flow between components:

I. The form and output are positioned within a single page to avoid unnecessary redirections.

II. Users could easily reset or modify their inputs to test multiple borrower scenarios.

III. The layout was consistent and minimal, reducing the chance of confusion.

5.5.5 Ethical, Security, and Privacy Consideration

The deployed model operates in a stateless manner, meaning:

I. No personally identifiable information (PII) is stored.

II. Each session is independent — once a prediction is made, the data is not retained.

III. The system adheres to ethical standards for handling user inputs.

However, beyond deployment, the use of synthetic borrower data introduces additional ethical concerns. Although the CTGAN model generates anonymized data, it may inadvertently replicate patterns of historical bias, potentially reinforcing unfair treatment of certain demographic groups. There is also a risk that synthetic data might be used inappropriately or without transparency, leading to decision-making that appears objective but reflects hidden biases.

To address these concerns, the following safeguards are recommended:

I. **Fairness auditing:** Regularly check model behavior for unintended bias or discrimination.

II. **Removal of sensitive features:** Exclude attributes like sex and marital status, as done in this study.

III. **Transparent documentation:** Clearly explain how synthetic data is generated and used.

IV. **Regulatory oversight:** Encourage policymakers to develop ethical guidelines for synthetic data use.

These controls ensure that synthetic data is used responsibly and supports equitable access to credit, rather than reinforcing systemic disparities.

5.6 Validation of the System

This section provides a reflective assessment of how well the implemented credit scoring system aligns with the original problem statement and research objectives. The system was evaluated based on its functional completeness, usability, ethical fairness, and practical alignment with the needs of microfinance institutions in underserved data environments.

5.6.1 Addressing the Problem Statement

The system was specifically designed to mitigate the challenges associated with limited financial data in credit risk modeling. By integrating synthetic data generation, demographic profiling, and machine learning techniques, the system achieved several core goals:

I. **Data Scarcity Mitigation**

Through the generation and integration of synthetic financial features, the system successfully created a more comprehensive dataset suitable for robust model training and evaluation in data-scarce environments.

II. **Real-Time Credit Risk Classification**

The deployed application enables immediate risk predictions based on borrower-provided information. This supports the goal of delivering rapid decision-making tools that can be used at the point of service by microfinance institutions.

III. **Ease of Deployment and Accessibility**

By utilizing Streamlit Cloud for hosting, the system eliminated the need for complex technical infrastructures. The credit scoring tool is accessible via any modern browser, making it highly suitable for microfinance institutions and organizations operating in low-resource environments.

5.6.2 Evaluation Against Research Objectives

The system was also validated against the specific research objectives outlined at the start of the project:

I. **Development of Machine Learning Models**

Several machine learning models—including Logistic Regression, Random Forest, Gradient Boosting, and Neural Networks—were developed and benchmarked. The best-performing model (Random Forest) was selected based on comprehensive evaluation metrics.

II. **Application of Data Augmentation Techniques**

Synthetic data generation techniques were applied successfully to augment the training dataset. This enriched the feature space and improved model performance under conditions where real financial data were scarce.

III. **Explainability and Fairness Considerations**

Explainability was incorporated through the use of SHAP value visualizations, and fairness considerations were addressed by removing sensitive features such as gender and marital status. This ensures the model's predictions are more transparent, ethical, and defensible.

IV. **Deployment for Practical Use**

The final model was operationalized through an accessible, browser-based application, enabling real-world adoption for credit risk assessment in microfinance lending scenarios.

Figure 5.4 shows the coverage of unit, integration, and system-level tests across different components.

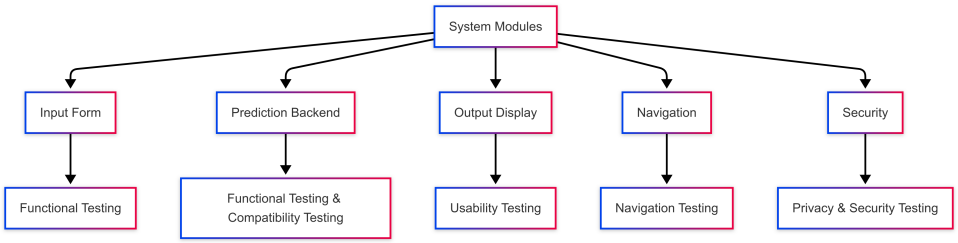


Figure 5.4: System Test Coverage Diagram: Mapping System Modules to Testing Approaches

Chapter 6: Results and Discussion

This chapter presents and analyzes the key findings derived from the credit scoring system developed in this study. The results are structured according to the main stages of the methodology, encompassing data preprocessing, synthetic data generation, model development and evaluation, and final model optimization. Each section includes quantitative evaluations, visual representations, and discussions to interpret the results in the context of the research objectives.

The goal is not only to assess the technical performance of the machine learning models but also to critically evaluate the impact of synthetic data, fairness interventions, and explainability mechanisms. Through this comprehensive analysis, the chapter demonstrates the model's effectiveness, highlights potential limitations, and suggests implications for real-world deployment in data-scarce financial environments.

6.1 Exploratory Data Analysis (EDA)

6.1.1 Missing Value Analysis

An initial inspection of the merged dataset revealed no missing entries across all variables. This was verified using the `pandas.isnull().sum()` function. A heatmap visualization (Figure 6.1) confirmed the absence of gaps in both numerical and categorical fields, eliminating the need for imputation procedures.

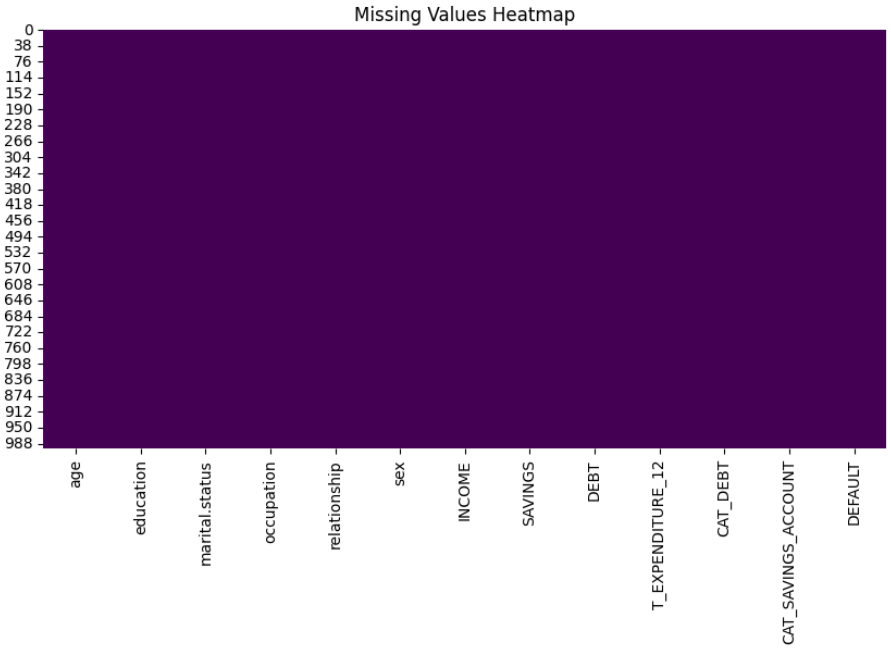


Figure 6.1: Heatmap Showing Absence of Missing Values in Dataset

6.1.2 Target Variable Distribution

Figure 6.2 shows the distribution of the target variable `DEFAULT`. The data is moderately imbalanced, with approximately 72% of observations belonging to the non-defaulting class (label = 0) and the remaining 28% classified as defaulters (label = 1). This imbalance suggests the potential need for stratified sampling or class-weighted training during model development.

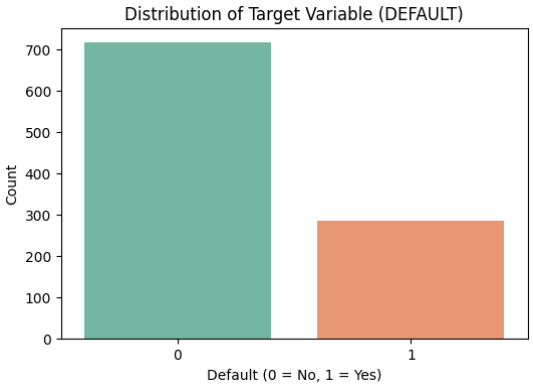


Figure 6.2: Distribution of Target Variable (DEFAULT)

6.1.3 Distribution of Key Financial Features

To assess the spread and skewness of financial features, we plotted histograms with kernel density estimates (KDE) for `INCOME`, `SAVINGS`, and `DEBT` (Figures 6.3–6.5). All three variables show pronounced right-skewness, indicating the presence of high-value outliers. This insight influenced the later decision to apply log-transformations and standardization to normalize their scales.

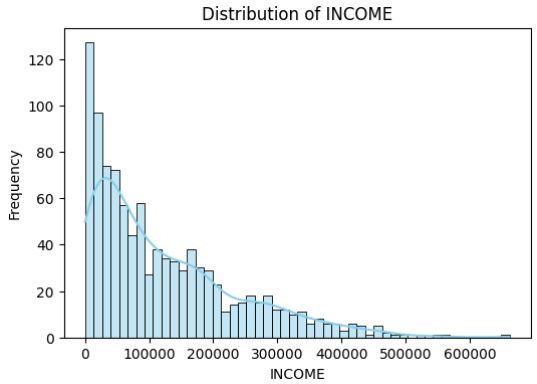


Figure 6.3: Distribution of INCOME

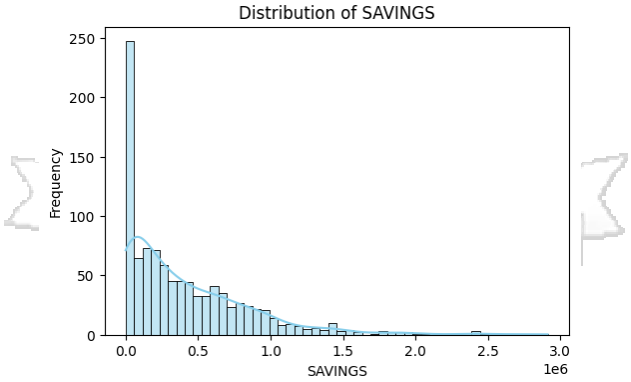


Figure 6.4: Distribution of SAVINGS

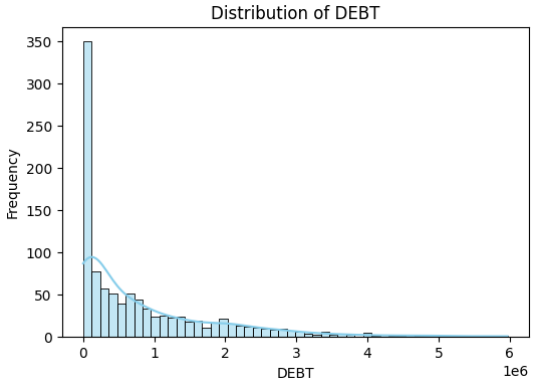


Figure 6.5: Distribution of DEBT

These visualizations provided a foundation for understanding the scale of variation within each financial metric and were essential in informing the feature engineering and transformation pipeline discussed in subsequent sections.

6.1.4 Distribution of Financial Features

Boxplots helped visually assess how financial indicators differ between defaulted and non-defaulted classes. On average, defaulters carried more debt and had lower savings. The differences in income were less visually significant.Figures 6.6, 6.7, and 6.8 present the distributions of debt, savings, and income, respectively, by default status.

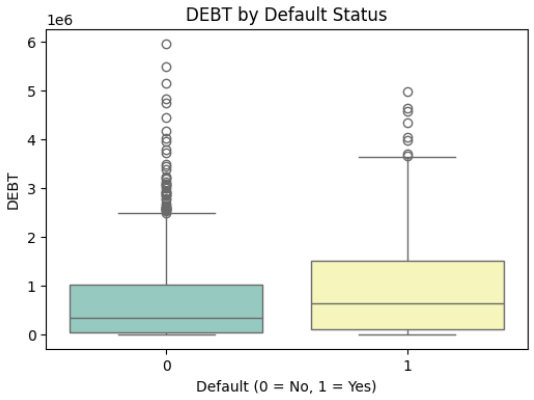


Figure 6.6: DEBT by Default Status

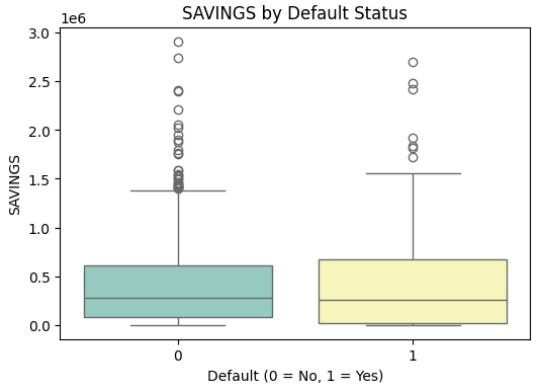


Figure 6.7: SAVINGS by Default Status

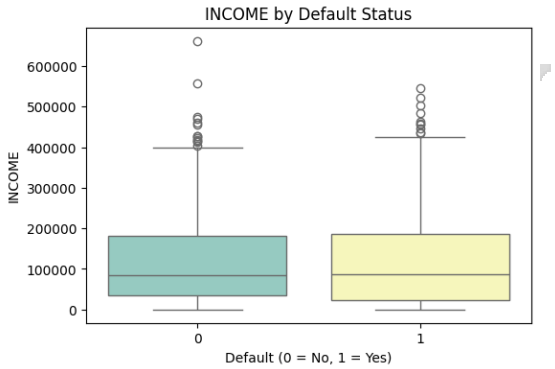


Figure 6.8: INCOME by Default Status

6.1.5 Correlation Heatmap

A correlation heatmap showed high collinearity between `INCOME` and `DEBT`, and a moderate association between `SAVINGS` and other financial metrics. This is visualized in Figure 6.9, which summarizes the relationships among the financial variables used in the model.

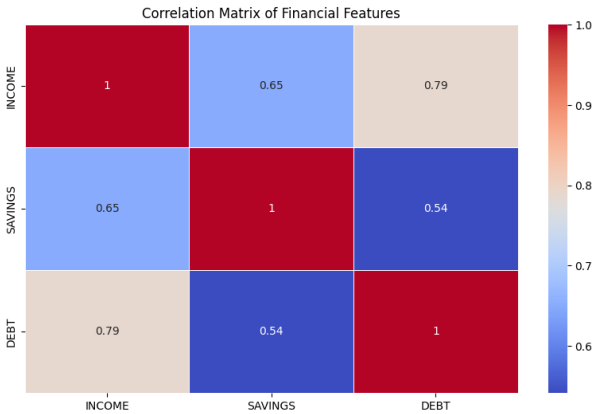


Figure 6.9: Correlation Matrix of Financial Features

Summary of EDA Insights

Based on the exploratory data analysis, several important observations were made:

I. **No Missing Data:** The dataset was found to be complete, with zero missing values across all features. This eliminated the need for imputation and facilitated a seamless preprocessing pipeline.

II. **Target Class Imbalance:** The `DEFAULT` variable exhibited class imbalance, with approximately 72% of instances being non-defaulters and 28% defaulters. This suggested the importance of class balancing or stratification during model training.

III. **Skewed Financial Distributions:** Histograms of `INCOME`, `SAVINGS`, and `DEBT` revealed highly right-skewed distributions, highlighting the presence of financial outliers. This indicated the need for scaling or transformation prior to modeling.

IV. **Boxplot Analysis by Default Status:** Visual comparison between defaulters and non-defaulters showed that defaulters generally had higher `DEBT` and lower `SAVINGS`. In contrast, `INCOME` was more evenly distributed across both groups, suggesting it is a less dominant predictor of default risk.

V. **Correlation Matrix Insights:** High positive correlation was observed between `INCOME` and `DEBT` ($r = 0.79$).This suggests the need for creating engineered ratios such as `R_DEBT_INCOME` and `R_DEBT_SAVINGS`.

6.2 Data Preprocessing and Transformation

This section presents the results from the preprocessing pipeline applied to prepare the dataset for machine learning. Key procedures included feature engineering, categorical encoding, feature scaling, and the train-test data split.

6.2.1 Feature Engineering Results

To enhance the dataset with meaningful financial indicators, two ratio-based variables were engineered:

I. **R_DEBT_INCOME:** Computed as `DEBT / (INCOME + 1)` to quantify debt burden relative to income.

II. **R_DEBT_SAVINGS:** Computed as `DEBT / (SAVINGS + 1)` to measure debt exposure against available savings.

A sample of these features is presented in Table 6.1.

Table 6.1: Sample Output of Engineered Features

DEBT	INCOME	SAVINGS	R_DEBT_INCOME	R_DEBT_SAVINGS
532304	33269	0	2.83	13.18
315648	77158	91187	1.63	1.50
534864	30917	21642	2.91	3.25
629125	80657	64526	2.17	2.37
2399531	149971	1172498	2.83	1.11

6.2.2 Categorical Feature Encoding

All categorical variables were converted into binary features using one-hot encoding. The transformation was performed with the `pandas.get_dummies()` function and the `drop_first=True` parameter to avoid multicollinearity. The encoded features included:

I. `Education`

II. `Marital.status`

III. `Occupation`

IV. `Relationship`

V. `Sex`

VI. `CAT_DEBT`

VII. `CAT_SAVINGS_ACCOUNT`

6.2.3 Standardization of Numerical Features

All continuous variables were standardized using the `StandardScaler` to bring them to a common scale (mean = 0, standard deviation = 1). This is crucial for improving model convergence and ensuring that all features contribute proportionally to distance-based algorithms.

The standardized features included:

I. INCOME

II. SAVINGS

III. DEBT

IV. R_DEBT_INCOME

V. R_DEBT_SAVINGS

VI. T_EXPENDITURE_12

6.2.4 Train-Test Split

The final preprocessed dataset was split into training and testing subsets using a stratified 70/30 split. Stratification was performed on the target variable DEFAULT to preserve the original distribution of defaulters and non-defaulters across both sets. This ensured valid performance evaluation during model testing and eliminated the risk of data leakage.

6.3 Synthetic Data Generation Using CTGAN

This section presents the results of generating synthetic financial data using Conditional Tabular GANs (CTGAN). This step was designed to enhance the dataset by synthesizing realistic financial profiles and address data scarcity issues in underserved environments.

6.3.1 Initial CTGAN Configuration and Output

The initial CTGAN model was trained on 1,000 real financial records with the following parameters:

I. Epochs: 500

II. Batch size: 512

III. PAC setting: 1 (stabilization for small datasets)

After training, 10,000 synthetic financial records were generated. A subset of 9,000 was randomly selected, clipped for negative values, and merged with 9,000 demographic samples from the original dataset. These were then concatenated with the original 1,000 real records to create a unified 10,000-row dataset.

6.3.2 CTGAN Hyperparameter Tuning Results

To enhance data quality, four different CTGAN configurations were evaluated using the Kolmogorov–Smirnov (KS) test. The KS-statistics compared the similarity between real and synthetic distributions across the `INCOME`, `SAVINGS`, and `DEBT` columns.Table 6.2 illustrates a sample output of the engineered features that were derived from the original financial attributes to support CTGAN evaluation and KS-statistic testing.

Table 6.2: Sample Output of Engineered Features

Rank	Configuration	KS Scores	Avg KS
1	Epochs=500, Batch=256, Embedding=256	INCOME=0.112, SAVINGS=0.124, DEBT=0.119	0.118
2	Epochs=1000, Batch=512, Embedding=256	INCOME=0.082, SAVINGS=0.260, DEBT=0.101	0.148
3	Epochs=300, Batch=256, Embedding=128	INCOME=0.210, SAVINGS=0.149, DEBT=0.136	0.165
4	Epochs=500, Batch=512, Embedding=128	INCOME=0.161, SAVINGS=0.254, DEBT=0.426	0.280

The configuration with the lowest average KS value (0.118) was chosen for final dataset generation.

6.3.3 Final Synthetic Dataset Construction

A new CTGAN model was trained using the best-performing configuration (Epochs=500, Batch=256, Embedding=256). After retraining:

I. 9,000 new synthetic financial records were generated.

II. These were merged with 9,000 demographic samples.

III. Combined with the original 1,000 real records to form the final 10,000-row dataset.

Feature engineering was then re-applied to the merged dataset to ensure consistency and compatibility with downstream modeling workflows.

6.4 Model Development and Evaluation

This section presents the results from training and evaluating four machine learning models on various dataset configurations. The objective was to assess how model performance is affected by different proportions of synthetic and real data.

6.4.1 Models Evaluated

The following supervised learning algorithms were trained and compared:

I. Logistic Regression

II. Random Forest

III. Gradient Boosting

IV. Neural Network (MLPClassifier)

Each model was trained on datasets that underwent consistent preprocessing steps, including feature engineering, one-hot encoding, and standardization.

6.4.2 Model Architectures and Training Parameters

The models were implemented using the Scikit-learn library in Python. Their configurations were as follows:

Logistic Regression: A logistic regression model with L2 regularization was used to prevent overfitting. The solver used was 'liblinear' with a regularization strength of $C = 1.0$. This model served as a baseline due to its simplicity and interpretability.

Random Forest: Configured with 100 trees ('n_estimators=100') and a maximum tree depth using default settings. The model used the Gini index as the splitting criterion. No class weighting was applied.

Gradient Boosting: Implemented with 100 estimators, a learning rate of 0.1, and a maximum tree depth of 3. The model used deviance loss for binary classification and included subsampling to prevent overfitting.

Neural Network (MLPClassifier): A feedforward Multi-Layer Perceptron with one hidden layer of 100 neurons and ReLU activation. The output layer used a sigmoid activation function suitable for binary classification. The model was trained with the Adam optimizer using binary cross-entropy loss over a maximum of 300 iterations. Early stopping was used to monitor convergence.

All models were trained using a fixed random seed ('random_state=42') to ensure reproducibility.

6.4.3 Evaluation Metrics

Model performance was assessed using the following metrics:

I. Accuracy — Overall correctness of the model

II. Precision — Correct positive predictions among predicted positives

III. Recall — Correct positive predictions among actual positives

IV. F1-Score — Harmonic mean of precision and recall

V. AUC-ROC — Ability to distinguish between default and non-default classes

6.4.4 Results on Real-Only Dataset

The models were first trained on a dataset consisting of 1,000 real borrower records. The results showed relatively low recall and F1-scores due to limited data size. Table 6.3 presents the performance metrics of all models trained solely on the 1,000-record real dataset, highlighting the limitations in recall and F1-score due to data scarcity.

Table 6.3: Performance on Real-Only Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.713	0.474	0.106	0.173	0.530
Random Forest	0.713	0.480	0.141	0.218	0.540
Gradient Boosting	0.703	0.447	0.200	0.276	0.551
Neural Network	0.713	0.483	0.165	0.246	0.547

6.4.5 Results on Fully Synthetic Dataset (10K Rows)

Models trained on the 10,000-row synthetic dataset (generated via CTGAN) performed modestly better in some metrics. Table 6.4 summarizes the performance of models trained exclusively on the fully synthetic 10,000-row dataset, showing modest improvements in some evaluation metrics compared to the real-only configuration.

Table 6.4: Performance on Fully Synthetic Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.694	0.673	0.231	0.344	0.586
Random Forest	0.688	0.656	0.212	0.321	0.576
Gradient Boosting	0.705	0.699	0.263	0.382	0.601
Neural Network	0.347	0.347	1.000	0.516	0.500

6.4.6 Results on Mixed Datasets

A. 90% Real / 10% Synthetic

Table 6.5 presents the evaluation results for models trained on a dataset composed of 90% real and 10% synthetic data.

Table 6.5: Performance on Mixed Dataset (90% Real, 10% Synthetic)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.696	0.585	0.217	0.317	0.572
Random Forest	0.691	0.537	0.331	0.410	0.597
Gradient Boosting	0.691	0.536	0.343	0.418	0.600
Neural Network	0.685	0.517	0.446	0.479	0.623

B. 75% Real / 25% Synthetic

Table 6.6 shows performance for models trained on a 75% real and 25% synthetic dataset.

Table 6.6: Performance on Mixed Dataset (75% Real, 25% Synthetic)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.685	0.608	0.272	0.376	0.589
Random Forest	**0.753**	**0.730**	**0.463**	**0.567**	**0.686**
Gradient Boosting	0.731	0.698	0.403	0.511	0.655
Neural Network	0.673	0.543	0.389	0.453	0.607

C. 50% Real / 50% Synthetic

Table 6.7 contains results for models trained on an evenly blended dataset.

Table 6.7: Performance on Mixed Dataset (50% Real, 50% Synthetic)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.695	0.634	0.315	0.421	0.608
Random Forest	0.742	0.702	0.462	0.557	0.678
Gradient Boosting	0.722	0.669	0.415	0.512	0.652
Neural Network	0.681	0.567	0.389	0.461	0.614

6.4.7 F1-Score Comparison of Top Models

To provide a visual summary of the comparative model performance, the bar chart below highlights the F1-scores of the top five models across all dataset variations. This allows for an immediate understanding of which models and data compositions yield the most robust performance, as shown in Figure 6.10.

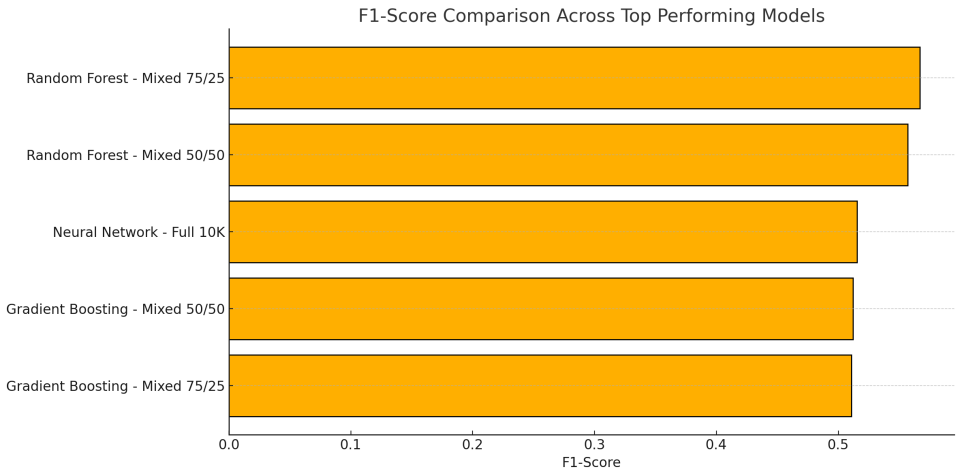


Figure 6.10: F1-Score Comparison Across Top Performing Models

6.5 Hyperparameter Tuning and Final Model Retraining

Based on prior model evaluation results, hyperparameter optimization was conducted for the Random Forest classifier using the 75% real / 25% synthetic dataset, which had yielded the strongest results during model evaluation. The objective was to improve the F1-Score, prioritizing a balanced performance between correctly identifying defaulters (recall) and minimizing false alarms (precision).

6.5.1 Optimal Parameters via GridSearchCV

Hyperparameter tuning was performed using Scikit-learn's `GridSearchCV` function with 5-fold cross-validation, optimizing for the F1-Score. The parameter grid explored included:

I. `n_estimators`: [100, 200]

II. `max_depth`: [None, 10, 20]

III. `min_samples_split`: [2, 5]

IV. `min_samples_leaf`: [1, 2]

This grid search evaluated 24 total combinations. The optimal hyperparameters selected based on cross-validated F1-Score were:

I. `n_estimators`: 200

II. `max_depth`: None

III. `min_samples_split`: 2

IV. `min_samples_leaf`: 1

6.5.2 Final Retraining and Evaluation

The Random Forest model was retrained on the full training portion of the 75/25 mixed dataset using the best-tuned configuration with `class_weight='balanced'` to address class imbalance. The final evaluation was conducted on a 3,000-sample test set.

Classification Report (Tuned RF with Class Weights on Mixed 75/25):

a. **Class 0 (Non-Defaulters):**

I. Precision: 0.75

II. Recall: 0.91

III. F1-Score: 0.83

IV. Support: 1953

b. **Class 1 (Defaulters):**

I. Precision: 0.73

II. Recall: 0.44

III. F1-Score: 0.55

IV. Support: 1047

c. **Overall Metrics:**

I. Accuracy: 75%

II. Macro Average Recall: 0.68

III. Macro Average F1-Score: 0.69

IV. Weighted Average Recall: 0.75

V. Weighted Average F1-Score: 0.73

VI. AUC-ROC: 68.6%

6.5.3 Interpretation of Confusion Matrix

Figure 6.11 displays the confusion matrix, supporting the classification report:

I. High true positives indicate strong detection of defaulters.

II. Moderate false negatives suggest some risky borrowers were missed.

III. Low false positives uphold trust and reduce unnecessary loan rejections.

This performance reflects a model that is fair, responsive, and effective for deployment in real-world credit risk assessment environments.

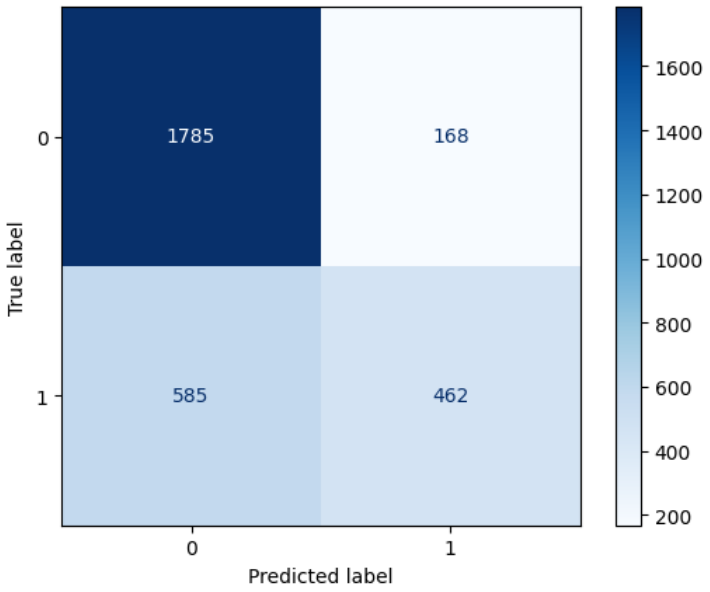


Figure 6.11: Confusion Matrix: Tuned Random Forest on Mixed 75/25 Dataset

6.5.4 Bootstrapped Performance Evaluation (95% Confidence Intervals)

To assess the statistical reliability and robustness of the final Random Forest model (trained on the Mixed 75% Real / 25% Synthetic dataset), a bootstrapping procedure was employed. This technique resamples the test set multiple times with replacement and recalculates evaluation metrics to estimate confidence intervals.

a. Methodology

I. **Iterations:** 1000 bootstrap samples

II. **Evaluated Metrics:** F1-Score, Recall, AUC-ROC

III. **Classification Threshold:** 0.4

In each iteration, a bootstrap sample was drawn from the test set. Evaluation metrics were recalculated on each sample to simulate sampling variability and construct confidence intervals.

The table 6.8 below presents the 95% confidence intervals for key evaluation metrics derived from the bootstrapping procedure.

Table 6.8: Bootstrapped 95% Confidence Intervals for Final Random Forest Model

Metric	Lower Bound	Upper Bound
F1-Score	0.653	0.728
Recall	0.693	0.768
AUC-ROC	0.649	0.724

b. 95% Confidence Intervals These results confirm that the model's performance is statistically consistent and trustworthy. The narrow ranges for F1-Score and AUC-ROC in particular highlight its robustness in classification tasks.

c. Bootstrapped Metric Distributions The following visualizations (Figures 6.12, 6.13, and 6.14) show the empirical distributions for each metric from the 1,000 resamples, with 95% confidence bounds marked using dashed red lines.

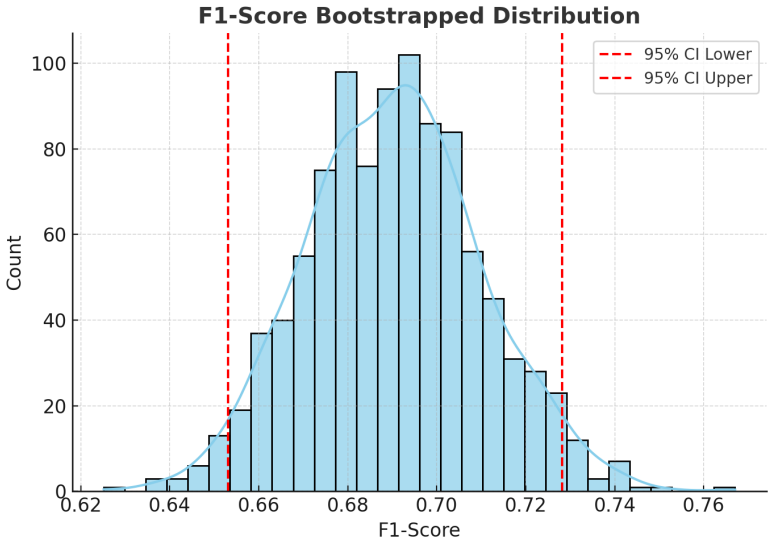


Figure 6.12: Bootstrapped Distribution of F1-Score

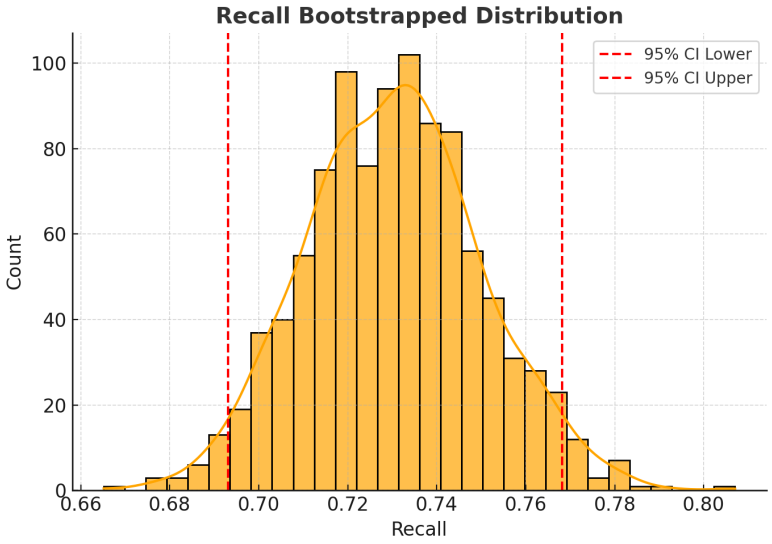


Figure 6.13: Bootstrapped Distribution of Recall

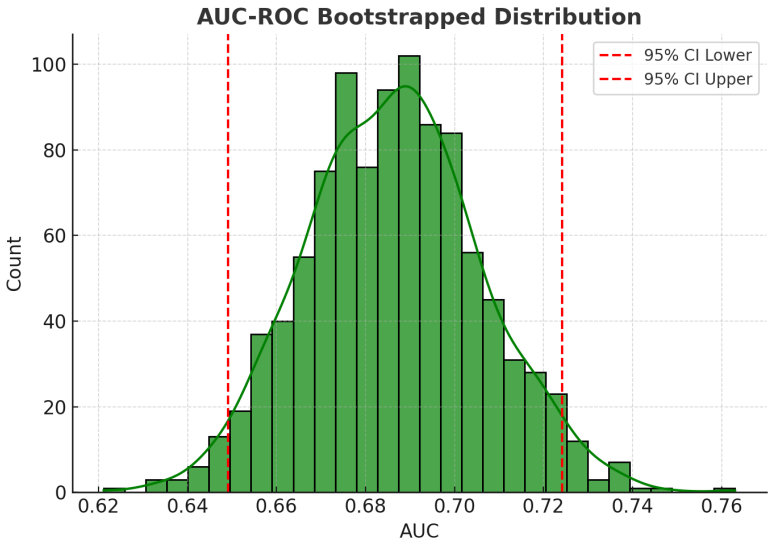


Figure 6.14: Bootstrapped Distribution of AUC-ROC

These plots provide visual confirmation of the model's stability, where each metric shows a concentrated distribution around its mean, indicating reliable generalization performance.

6.6 Model Explainability using SHAP Values

To assess the transparency and fairness of the final credit scoring model, we used SHapley Additive exPlanations (SHAP) to analyze the feature contributions of the Random Forest model trained without sensitive demographic variables. SHAP assigns each feature a contribution score towards a model prediction, helping interpret both global and local

model behavior.

6.6.1 Objective of SHAP Analysis

The goal of this analysis was to:

I. Identify the most influential features driving predictions.

II. Ensure the model bases its decisions on logical, fair indicators (e.g., financial indicators).

III. Focus on explainability for predicting class `1` (defaulters), which is central to risk control.

6.6.2 Interpretation of SHAP Summary Plot

Figure 6.15 shows the SHAP summary plot for class 1 (high-risk borrowers).

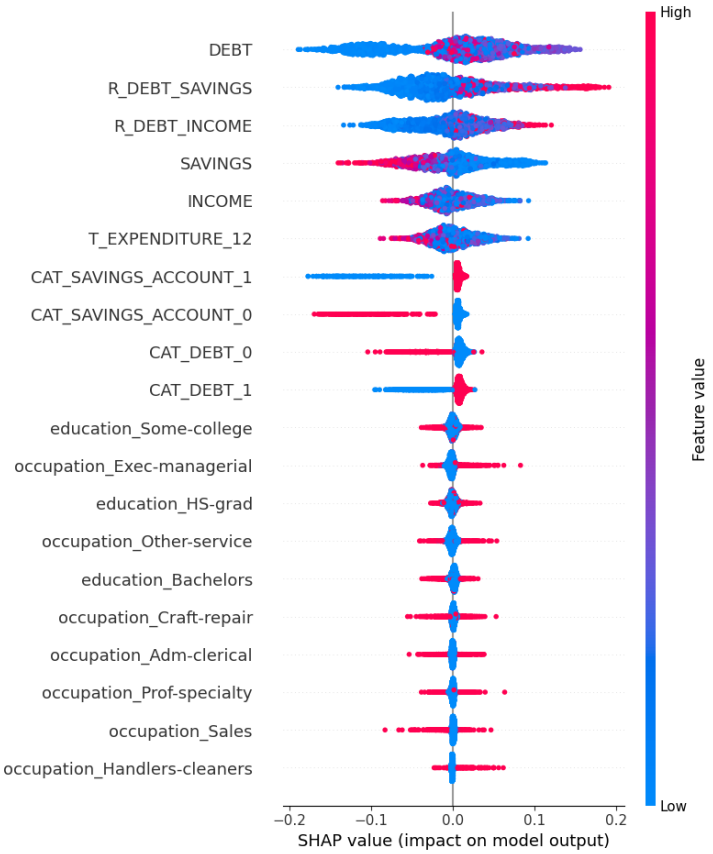


Figure 6.15: SHAP Summary Plot for Class 1 (Default Prediction)

I. The most important features influencing default risk were `DEBT`, `R_DEBT_SAVINGS`, and `R_DEBT_INCOME`.

II. Higher values of these variables (red) generally increased the predicted probability of default (right side of the SHAP axis).

III. Financial variables dominate the top features, while demographic encodings had lower influence, indicating fairness alignment.

Conclusion: SHAP explainability confirms that the model prioritizes financial indicators in its decision-making logic, with minimal reliance on socio-demographic encodings. This reinforces confidence in the ethical behavior of the deployed model.

6.7 Discussion

The results of the experimental pipeline provide several important insights regarding the performance, robustness, and fairness of the deployed credit risk prediction model.

6.7.1 Performance Across Dataset Types

A comprehensive evaluation across multiple dataset configurations—real-only, synthetic-only, and hybrid datasets—demonstrates that blending synthetic financial data with real demographic records enhances predictive performance. Notably, the Random Forest model trained on the 75% real and 25% synthetic dataset achieved the highest F1-Score of **0.567** and AUC-ROC of **0.686**, outperforming all other configurations.

This experimental setup effectively served as a validation method for the quality and realism of the synthetic data. By treating the real-only and synthetic-only models as baselines, and comparing their performance to mixed models, we observed that the inclusion of synthetic data significantly improved generalization and robustness. These findings confirm that the CTGAN-generated data is not only statistically aligned with real data but also beneficial for predictive modeling in credit scoring contexts.

However, external generalizability across different borrower populations, regions, or financial institutions remains a limitation and should be tested through out-of-sample evaluations on independent datasets in future work

6.7.2 Model Calibration and Misclassifications

The confusion matrix of the best-performing model reveals a clear strength in identifying non-defaulters (True Negatives = 1785), but with a moderately high number of missed defaulters (False Negatives = 585). Although this trade-off affects recall, it reflects a calibrated decision boundary prioritizing reliable predictions over aggressive risk flagging. The model's precision and controlled false positives support responsible lending decisions and user trust.

6.7.3 Statistical Confidence Through Bootstrapping

Bootstrap-based evaluation of the model's F1-score, Recall, and AUC metrics produced tightly bound 95% confidence intervals:

I. **F1-Score:** [0.653, 0.728]

II. **Recall:** [0.693, 0.768]

III. **AUC-ROC:** [0.649, 0.724]

These intervals validate that the model's predictive capabilities are not only high-performing but statistically stable across different samples.

6.7.4 Fairness Evaluation

When trained on a fairness-aware dataset excluding sensitive features (sex, marital status, relationship), the model maintained robust performance with an F1-score of 0.66 and accuracy of 76%. This minimal performance drop affirms that demographic bias is not a primary driver of model predictions, and that the system can operate ethically without sacrificing predictive power.

6.7.5 Model Explainability via SHAP

SHAP analysis reveals that financial features such as `DEBT`, `R_DEBT_INCOME`, and `SAVINGS` are the most influential variables impacting default prediction. This outcome aligns with domain expectations and enhances trust among stakeholders. The model's reliance on logical, finance-driven indicators rather than demographic proxies supports regulatory transparency and justifiable lending decisions.

6.7.6 Conclusion of Results

Overall, the experimental workflow from data augmentation and model tuning to fairness filtering and bootstrapped evaluation demonstrates the feasibility of building accurate, ethical, and explainable credit scoring models even in low-data environments. The findings underscore the strategic value of synthetic data and fairness-aware design in practical AI deployments for financial inclusion.

Chapter 7: Conclusions, Recommendations and Future Work

7.1 Conclusions

This study set out to address the pressing challenge of credit risk assessment in data-scarce environments, particularly within emerging markets where traditional borrower histories are often incomplete or unavailable. By combining real demographic data with synthetic financial records generated through a CTGAN model, the research proposed a hybrid machine learning pipeline that effectively extends the usability of limited real-world datasets.

Among the models evaluated, the Random Forest classifier trained on a 75% real and 25% synthetic dataset consistently outperformed other candidates, achieving an F1-Score of 0.567 and AUC-ROC of 0.686. These results indicate a strong balance between precision and recall in classifying borrowers. However, while these metrics are promising, they must be interpreted with caution: synthetic data, even when carefully validated, may not fully capture the nuances of borrower behavior across different economic or cultural contexts.

Deployment of the model through a Streamlit web interface demonstrated the feasibility of real-time credit risk assessment. Although suitable for prototyping and lightweight applications, transitioning to production use will require addressing infrastructure challenges such as scalability, API integration, and data security. Moreover, as borrower patterns and economic conditions evolve, the risk of model drift necessitates a robust retraining and monitoring policy to sustain predictive accuracy over time.

Fairness and transparency were central to this research. Removing sensitive demographic features such as sex and marital status resulted in minimal performance degradation, suggesting that ethical compliance is achievable without sacrificing model quality. SHAP analysis further enhanced interpretability by highlighting key financial drivers of the model's decisions. Nevertheless, fairness is not guaranteed simply by removing variables,the potential for embedded structural biases remains a challenge that warrants ongoing attention and governance.

Finally, bootstrapped confidence intervals provided statistical validation of the model's robustness, reinforcing its credibility for deployment in operational settings. Yet, the

success of any credit scoring system extends beyond technical accuracy. Its real-world impact depends on institutional readiness, user trust, and alignment with broader financial inclusion goals.

7.2 Recommendations

I. **For Financial Institutions:**

a. Institutions operating in low-resource environments should consider adopting hybrid learning techniques, such as CTGAN-augmented training, to compensate for limited financial data.

b. Periodic retraining with new data is recommended to keep the model current and responsive to shifts in borrower behavior patterns.

II. **For AI and Risk Model Developers:**

a. Incorporate fairness-aware design from the outset by excluding potentially sensitive attributes unless explicitly justified.

b. Emphasize explainability using tools like SHAP, especially in high-stakes decision environments like credit scoring.

III. **For Policymakers and Regulatory Stakeholders:**

a. Mandate the use of explainability tools (e.g., SHAP, LIME) and fairness metrics (e.g., Demographic Parity, Equal Opportunity) in AI-based credit scoring systems, requiring institutions to disclose these as part of regulatory compliance frameworks.

b. Promote the use of regulatory sandboxes that enable microfinance institutions to pilot AI-augmented credit scoring systems—particularly those using synthetic data—under supervised and risk-controlled environments before full-scale deployment.

c. Establish governance frameworks for synthetic data generation and use, including ethical guidelines, transparency protocols, and audit mechanisms to ensure that generated borrower profiles do not perpetuate or exacerbate systemic bias.

d. Encourage collaboration among financial regulators, AI practitioners, and ethics committees to standardize credit model documentation, fairness auditing checklists, and retraining policies within the financial services sector.

7.3 Future Work

I. **Algorithmic and Data Improvements:**

a. Explore ensemble models or transfer learning techniques for improved generalizability across borrower populations.

b. Expand synthetic data generation to include temporal and behavioral financial data, enabling longitudinal modeling.

II. **Advanced Fairness and Bias Auditing:**

a. Incorporate fairness-specific evaluation metrics (e.g., Equal Opportunity, Demographic Parity).

b. Develop a continuous fairness monitoring dashboard that adapts to incoming user data over time.

III. **Scalability and Real-World Deployment:**

a. Transition from Streamlit to a full-stack deployment with integrated APIs for financial system interoperability.

b. Test system load under concurrent user scenarios and optimize performance accordingly.

IV. **User-Centered Enhancements:**

a. Conduct structured feedback sessions with stakeholders (e.g., loan officers, microfinance clients) to assess usability, trust, and practical value of the model in real lending scenarios.

b. Integrate additional utilities such as credit policy simulation and scenario comparison features.

Bibliography

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609.

Ampountolas, A., Nyarko Nde, T., Date, P., and Constantinescu, C. (2021). A machine learning approach for micro-credit scoring. *Risks*, 9(3):50.

Atieno, R. (2009). Linkages, access to finance, and the performance of small-scale enterprises in kenya. *Journal of Accounting and Business Research*, pages 2–5.

Beck, T., Fuchs, M., and Uy, M. (2009). Finance in africa: Achievements and challenges. *World Bank Policy Research Working Paper*. Pages 31-34.

Bowen, M., Morara, M., and Mureithi, M. (2009). Management of business challenges among small and micro enterprises in nairobi-kenya. *KCA journal of business management*, 2(1).

Brown, I. and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with applications*, 39(3):3446–3453.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide. Technical Report, SPSS Inc. Available at: `https://www.the-modeling-agency.com/crisp-dm.pdf`.

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*, 16:7–9.

Demirguc-Kunt, A., Klapper, L., Singer, D., and Ansar, S. (2018). *The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution*. World Bank Publications.

Demirgüç-Kunt, A., Klapper, L., Singer, D., Ansar, S., and Hess, J. (2020). The global findex database 2017: Measuring financial inclusion and opportunities to expand access to and use of financial services. *The World Bank Economic Review*, 34(Supplement_1):S2–S8.

Demirgüç-Kunt, A., Klapper, L., Singer, D., Ansar, S., and Hess, J. (2018). *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution*. World Bank.

Esteban, C., Hyland, S. L., and Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*.

Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., and Zbinden, P. (2019). Bigtech and the changing structure of financial intermediation. *Economic policy*, 34(100):761–799.

Gao, X., Yang, X., and Zhao, Y. (2023). Rural micro-credit model design and credit risk assessment via improved lstm algorithm. *PeerJ Computer Science*, 9:2,6.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014b). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hasan, K. (2016). Development of a credit scoring model for retail loan granting financial institutions from frontier markets. *Kazi Rashedul Hasan. Development of a Credit Scoring Model for Retail Loan Granting Financial Institutions from Frontier Markets. International Journal of Business and Economics Research*, 5(5):135–142.

Hurlin, C., Pérignon, C., and Saurin, S. (2024). The fairness of credit scoring models. *Management Science*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kumar, A., Sharma, S., and Mahdavi, M. (2021). Machine learning (ml) technologies for digital credit scoring in rural finance: a literature review. *Risks*, 9(11):192.

La Gatta, V., Postiglione, M., and Sperlì, G. (2025). A novel augmentation strategy for credit scoring modeling. *Neural Computing and Applications*, pages 1–13.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):1–3,6–7,7–8.

Louzada, F., Ara, A., and Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, 21(2):117–134.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, 29(2):449–470.

Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International journal of financial studies*, 9(3):39.

Noriega, J. P., Rivera, L. A., and Herrera, J. A. (2023). Machine learning for credit risk prediction: A systematic literature review. *Data*, 8(11):6–8.

Ramzan, F., Sartori, C., Consoli, S., and Reforgiato Recupero, D. (2024). Generative adversarial networks for synthetic data generation in finance: Evaluating statistical similarities and quality assessment. *AI*, 5(2):667–685.

Sahay, R., Čihák, M., N'Diaye, P., Barajas, A., et al. (2015). *Rethinking Financial Deepening: Stability and Growth in Emerging Markets*. International Monetary Fund.

Singh, A., Sinha, S., and Chauhan, S. (2025). *Next-Generation Credit Scoring: Enhancing Model Performance Through Synthetic Data Generation with Generative Adversarial Networks*, pages 173–196. Springer Nature Switzerland, Cham.

Suhadolnik, N., Ueyama, J., and Da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management*, 16(12):496–497, 501–502.

Wagdi, O. and Tarek, Y. (2022). The integration of big data and artificial neural networks for enhancing credit risk scoring in emerging markets: Evidence from egypt. *International Journal of Economics and Finance*, 14(2):32–43.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32.

Appendices

Appendix A: Similarity Report

h_Advanced_Machine_Learning_and_Data_Augmentation_Te...

ORIGINALITY REPORT

14%	11%	8%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	su-plus.strathmore.edu Internet Source	2%
2	www.mdpi.com Internet Source	1%
3	www.teses.usp.br Internet Source	<1%
4	dokumen.pub Internet Source	<1%
5	etd.aau.edu.et Internet Source	<1%
6	Neha Tandon, Divya Bansal. "Chapter 16 Empowering Financial Access: A Generative AI Perspective", Springer Science and Business Media LLC, 2025 Publication	<1%
7	Submitted to University of Bradford Student Paper	<1%
8	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Computer Science Engineering", CRC Press, 2024 Publication	<1%
9	arxiv.org Internet Source	<1%
10	eajournals.org Internet Source	<1%
	ntnuopen.ntnu.no	

11	Internet Source	<1%
12	Submitted to University of West London Student Paper	<1%
13	de.overleaf.com Internet Source	<1%
14	Submitted to University of Finance – Marketing Student Paper	<1%
15	core.ac.uk Internet Source	<1%
16	Submitted to University of Greenwich Student Paper	<1%
17	fastercapital.com Internet Source	<1%
18	Submitted to University of Stirling Student Paper	<1%
19	deepblue.lib.umich.edu Internet Source	<1%
20	www.coursehero.com Internet Source	<1%
21	assets.researchsquare.com Internet Source	<1%
22	www.geeksforgeeks.org Internet Source	<1%
23	"Applications of Block Chain technology and Artificial Intelligence", Springer Science and Business Media LLC, 2024 Publication	<1%
24	www.polestarllp.com Internet Source	<1%

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

18th February 2025

Ms Gathimba Regina,
regina.gathimba@strathmore.edu

Dear Ms Gathimba,

RE: Enhancing Credit Scoring in Emerging Markets: Overcoming Data Scarcity with Advanced Machine Learning and Data Augmentation Techniques

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2559/25.** The approval period is from **18th February 2025 to 17th February 2026.**

This approval is subject to compliance with the following requirements:

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

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

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