A MACHINE LEARNING-BASED LOAN DEFAULT PREDICTION FOR INDIVIDUALS IN RURAL COOPERATIVE SOCIETIES

BY

ORJI FRANKLIN CHUKWUEBUKA

REG. NO: GOU/U22/CSC/826

PROJECT REPORT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AND MATHEMATICS, FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY, GODFREY OKOYE UNIVERSITY, ENUGU STATE

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR AWARDING THE DEGREE OF BACHELOR OF SCIENCE IN COMPUTER SCIENCE

SUPERVISOR:

ENGR. CAMILLUS NJOKU

MAY, 2026

APPROVAL PAGE

This research, titled "a Machine Learning-Based Loan Default Prediction for Individuals in Rural Cooperative Societies," has been assessed and approved by the Department of Computer Science Committees in the Faculty of computing and information technology, Godfrey Okoye University, Enugu, Nigeria.

Engr. Camillus Njoku	....	....
Supervisor	Signature	Date
External Examiner	....	....
External Examiner	Signature	Date
Dr. Frank Okebanama	....	....
HOD, Department of Computer Science and Mathematics	Signature	Date
Prof. I.A. Ajah	....	....
Dean, Faculty of Computing And Information Technology.	Signature	Date

CERTIFICATION

I certify that I completed the research included therein and that it was primarily based on my observations and investigation. Consequently, I will fully accept the University's decision if I am found to have cheated on the assignment.

ORJI FRANKLIN CHUKWUEBUKA

REG. NO: GOU/U22/CSC/826

DEDICATION

This project is dedicated to God Almighty, the source of all wisdom, knowledge, and strength, whose grace sustained me through every challenge and made the completion of this project possible.

ACKNOWLEDGMENTS

I give all glory, honor, and gratitude to Almighty God for His infinite mercy, wisdom, strength, and guidance throughout the course of my academic pursuit and for granting me the ability to successfully complete this project.

My profound appreciation goes to my project supervisor, Engr Camillus Njoku, whose invaluable guidance, constructive criticism, encouragement, and professional advice greatly contributed to the success of this research work. Your patience and willingness to share your knowledge made this project a rewarding learning experience.

I am equally grateful to the Head of Department, Dr. Frankline Okebanama, and all the lecturers in the Department for their dedication, support, and commitment to imparting knowledge throughout my years of study. The academic foundation and skills acquired under your tutelage have played a significant role in shaping my educational and professional growth.

I wish to express my heartfelt gratitude to my parents and family members for their unwavering love, prayers, sacrifices, financial support, and encouragement. Their belief in my abilities has been a constant source of motivation and strength throughout my academic journey.

My sincere appreciation also goes to my friends, classmates, and well-wishers who contributed directly or indirectly to the completion of this project through their advice, encouragement, and support. Your contributions, no matter how small, are deeply appreciated.

Finally, I acknowledge everyone who has assisted me in one way or another toward the successful completion of this project. May God richly bless you all.

Abstract

The issue of loan default remains a significant challenge among rural cooperatives due to a lack of suitable methods for assessing credit risk, the limited amount of information available about borrowers, and an over-reliance on manual methods to make lending decisions. These limitations often result in larger amounts of non-performing loans and associated losses. This study developed a machine learning-based loan default predictive system to help rural cooperatives make data driven decisions about the lending status of individuals applying for credit. The proposed system used the Extreme Gradient Boost (XGBoost) algorithm because of its demonstrated high performance on structured classification problems. The Home Credit dataset provided the borrower and loan data used in this study and the data were cleaned, feature selected, feature engineered, treated for missing values, and encoded prior to training the model. The trained model was incorporated into a web-based application that was built using Flask, React Native, and MongoDB, allowing for real-time assessment of loan risk. The model's performance was evaluated based

on accuracy, confusion matrix, and Receiver Operating Characteristic (ROC) analysis. The model demonstrated an accuracy of 85.4% with an AUC of 0.7569, thus providing evidence of good predictive capability despite class imbalance in the dataset. The system will provide rural cooperatives with an intelligent decision support tool that can assist in identifying high-risk borrowers prior to loan approval; therefore, reducing the risk of defaulting on loans, improving credit management, and promoting sustainable financial activities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

The banking sector is an important component or a substantial pillar of the growth of the economy of all countries in the world. It serves all the sections of society by providing financial services that are vital for their day-to-day operations and survival, and such services include deposits, withdrawals, and money lending (Hota et al., 2025). Different categories of banks exist in various areas based on the population density and the economic status of the particular place in order to cater to the services required. In general, there are classifications for banking institutions. Each classification is established to provide for the financial needs of different-size communities with different levels of economic development. The type of bank that exists in a given area depends on the characteristics of that area's economy and the needs of its businesses and consumers. Therefore, the banking sector consists of several classifications, which include public-sector banks, regional banks, urban cooperative banks, rural cooperative banks and foreign banks, with each of these classifications for the banks created to meet a specific financial need from a particular customer base.

Banks lend loans to every sector for their growth and expansion, following the country's overall development. For a bank to provide a loan, a trust or guarantee must have been built that the borrower will repay the amount, along with the agreed interest, and within the timeframe or duration agreed for the loan (Hota et al., 2025). A loan has three main parts: the principal (the amount borrowed), the interest rate, and the loan duration. When someone applies for a loan, the application goes through several checks before approval. The most important factor is the borrower's credit score, which shows if they are likely to repay on time. Lenders also look at income, job status or business turnover, debt-to-income ratio for self-employed people, collateral, and down payment history (Odegua, 2019). Other general factors include age, gender, marital status, family situation, and the borrower's relationship with the bank. By reviewing these details, lenders decide if someone is eligible for a loan. Today, banks especially focus on credit scores to identify potential defaulters, and loan default prediction is one of the most critical problems faced by banks and other financial institutions (Hota et al., 2025), as it has a significant impact on profits. Banking institutions have traditionally used criteria for assessing loan applications that are now being proven to be inadequate in predicting performance of the loan. The continuous increase in the number of non-performing loans is an indication that conventional methods have been ineffective in identifying high-risk borrowers, which is why there is a need for developing better and smarter models for evaluating credit risks.

In developing economies, especially within their rural communities, loan default rates are much higher as a result of factors like irregular income streams, limited financial literacy and insufficient mechanisms for credit-worthiness assessment. Rural borrowers in low-income settings are generally engaged in small-scale, subsistence-based economic activities, possess limited assets, and rely heavily on agriculture or informal microenterprises. This dependence makes them more susceptible to seasonal trends and climate change induced stresses like drought, floods, etc. and price fluctuations. Studies on the clients of microfinance institutions in rural areas reveal that they are often plagued by irregular cash flows, unstable and low earnings, and less formal education, which all limit their capacity to plan and manage their debt obligations appropriately (Sutariya & Arora, 2023). Further, these rural households have access to informal credit sources, and regularly move from one lender to another, highlighting liquidity needs and lack of access to mainstream banking systems. The fact that these observed qualities indicate that the credit worthiness of the rural borrowers cannot be reflected on orthodox banking indicators like payslips or formal collateral, suggests that rural credit users are not suitable for orthodox approaches to credit (Advans International, 2026).

Geographic dispersion, lack of infrastructure, and high operating expenses are factors that make it difficult to reach clients and monitor loans at financial institutions in rural areas, such as cooperative societies and microfinance organizations. There are higher costs per client for branches in less populated areas, and limited transport and digital access decreases the frequency and effectiveness of branch-borrower interaction (Patel & Tirthani, 2025). These institutions also are more risky given their portfolio of clients is largely drawn from the agriculture and informal sectors, which are highly susceptible to weather and market fluctuations. This can lead to a lot of clients defaulting at the same time during hard times. Furthermore, cooperative organizations also face challenges with weak management, ineffective internal controls, and political influence on lending, which makes the management of risks difficult (Ganawah et al., 2025).

One of the biggest challenges for rural cooperatives in giving out loans is the lack of good borrower data. This will make it difficult to determine the credit risk and utilize sophisticated risk analysis tools (Ramadevi et al., 2025). Paper records are still used in many cooperatives or their databases are not consistent or standardized for tracking repayments, monitoring performance, identifying borrowers with loans from multiple sources, etc. One major challenge with rural clients is that they lack documents such as official IDs, land titles and proof of regular incomes that limits the effectiveness of traditional credit scoring and collateral requirements (Advans International, 2026). Research indicates default risks are greater in rural areas due to unstable income, cash flow that is seasonal and lack of formal credit history. When earnings are irregular and income is low, families may also find themselves without savings or insurance, and they may not be able to pay their loans when unexpected events occur like a crop loss, drop in the market, or health crisis. Lenders lack sufficient credit information to assess risk, resulting in limited loan eligibility or increased interest rates, making loans more burdensome for borrowers (Patel & Tirthani, 2025). To do this, new approaches such as machine learning approaches to default should also take into account rural realities, such as income fluctuations, seasonal patterns and other indicators such as group guarantees, regular saving patterns and mobile money. This would make the risks of inconsistent incomes and inadequate credit information more apparent.

Machine learning is growing rapidly in financial decision-making and is changing how credit risk is assessed by using advanced pattern recognition and data-driven analytics (Lai, 2020). These approaches have been effective in uncovering more complex, non-linear relationships within large datasets that traditional models are unable to capture, thereby improving loan default prediction accuracy (Ochieng, 2025). In recent studies, the use of ML models such as Random Forest (RF), XGBoost and AdaBoost has been seen to be effective in predicting loan defaults with over 80% accuracy in various banking and microfinance contexts (Bhatore et al., 2022). XGBoost has achieved high accuracy (up to 83%) and low Brier scores in the context of credit risk management in microfinance institutions, outperforming logistic regression. It is particularly important in the case of rural cooperative societies with high default risks (Ochieng, 2025). Research conducted from 2020 to 2025 has emphasized the effectiveness of ML in dealing with imbalanced datasets and real-time decision-making, which can lead to a decrease in non-performing loans and promote financial inclusion in regions where it is lacking (Malhotra & Malhotra, 2024; Babarinde et al., 2025). While machine learning is gaining momentum in the field of credit risk assessment, there are still limited systems available that are practical, specific to the rural context and can predict loan defaults. These areas have a number of specific challenges due to data constraints and unique socio-economic factors.

1.2 Statement of the Problem

Lending is an important aspect of economic development; however, when abused, it becomes a risk for the financial system and economic growth as a whole. The current way of assessing the creditworthiness of rural residents is fundamentally flawed due to its inconsistencies. Most residents of rural, low-income communities borrow money from friends, family, and neighbors to provide for their families' basic needs, and engage in many different, typically small rather than larger-scale, subsistence-based economic activities; have little or no assets; rely almost exclusively on either agricultural or informal microenterprise-based businesses; and have various other forms of subsistence income.

One major impediment to rural cooperative lending institutions is the lack of reliable information regarding borrowers; as a result, rural lending institutions find it more difficult to assess borrowers' credit risk and apply the latest and most sophisticated risk analysis methodologies. In general terms, there remains significant room for improvement regarding machine learning models used to predict credit risk; however, there are a number of issues related to data quality, class imbalance, model interpretability, data privacy concerns, and changing economic conditions that need to be addressed. Many of these problems are exacerbated when it comes to lending in rural cooperatives, as rural borrowers are often unable to produce either complete formal or informal financial documents for use in assessing their credit risk.

1.3 Aim and Objectives

The aim of this study is to develop a Machine Learning-Based Loan Default Prediction for Individuals in Rural Cooperative Societies. The objectives are to:

1. Identify key socio-economic and financial factors that influence loan default in rural communities.

2. Collect, preprocess, and analyse relevant borrower and loan data for effective predictive modelling.

3. Train and evaluate different predictive algorithms to determine the most accurate model for loan default prediction, using appropriate performance metrics.

4. Integrate the model into a web-based software application system that enables real-time validation and supports rural lenders in assessing borrower credit risk before loan approval.

1.4 Significance of the Study

Loan default is an issue faced by small rural banks, cooperatives and microfinance institutions. A lack of credit history for borrowers and poor risk evaluation methods, as well as concerns with the overall economy are all contributing factors. The development of a machine-learning-based system for predicting loan defaults will assist rural financial institutions in lending decisions, minimizing financial losses and aiding in the long-term development of rural credit programs. Ultimately, the development of such a system would provide lenders a means to improve their ability to manage credit risk and reduce the potential for future defaults by providing information regarding potential defaulters before to approving loans. There are multiple stakeholders who would benefit from such a study, such as:

1. Rural Financial Institutions: Will have the capability to identify potential high-risk borrowers, thus decreasing the risk of loan defaults.

2. Cooperative Societies and Microfinance Institutions: Will allow for an improvement in the accuracy of credit assessments and enable these organizations to provide loans to borrowers in a sustainable manner.

3. Borrowers in Rural Communities: Will enhance the fairness and impartiality of credit evaluations based upon data as opposed to opinion.

4. Policymakers and Development Agencies: The results of this study can assist in the establishment of better rural financial policies and in the establishment of sound credit risk management frameworks.

5. Researchers and Academia: Add to the existing knowledge about the use of machine learning for the prediction of financial risks.

1.5 Scope of the study

The study will focus on developing a machine learning based loan default prediction system tasked with assisting rural cooperative banks in evaluating borrower credit risk. The study will collect, preprocess and analyse appropriate borrower and loan-related data obtained from relevant datasets to train multiple machine learning algorithms and adopt the one with the best performance metrics. The research also covers the development of a web-based software prediction system that can check loan applications in real time. Lenders will be able to enter borrower details and get predictions about the chance of loan default.

1.6 Operational Definition of Terms

1. **Credit Risk:** The possibility that a borrower may fail to repay a loan or meet contractual obligations, potentially causing financial loss to the lending institution.
2. **Financial Factors**: Borrower-specific monetary indicators, including previous loans, repayment history, debt-to-income ratio, and savings, are used as predictors in the model.
3. **Loan Default:** Failure of a borrower to meet scheduled loan repayments within the agreed period, including principal and interest.

4. **Machine Learning (ML)** A branch of artificial intelligence that enables systems to learn patterns from historical data and make predictions or decisions without having to be explicitly programmed.
5. **Predictive Modelling:** The process of using historical and current data to create a model capable of forecasting future outcomes, such as the likelihood of loan default.
6. **Socio-Economic Factors**: Characteristics of borrowers, such as income, occupation, education, family size, and land ownership, which may influence their ability to repay loans.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter summarizes all the related works and systems related to the proposed system. It discusses the theoretical underpinnings of loan default and risk prediction systems, conceptual frameworks, the use of machine learning, and the current trends. Moreover, related literature discussion points out successful and unsuccessful work done till now and then the need of localised intelligent loan prediction system in rural cooperative society is identified.

2.2 Theoretical Background

This section provides an explanation about the theories that are used to build a machine learning system to predict the risk of loan default for rural cooperative societies. It explores important concepts and frameworks related to credit risk, borrower behavior and the application of data to inform financial decision-making. These theories illustrate how borrower characteristics, financial information and loan history can be applied to predict the likelihood of default. Knowing these concepts helps in improving the accuracy and efficiency of credit risk prediction in rural cooperative lending through machine learning.

2.2.1 Loan Default

In the event that the debtor fails to comply with the legal obligation as stipulated in a debt contract, for example, failure to send the scheduled payments or failure to adhere to the loan covenants, default occurs (Azeez & Ajayi, 2023). A default is a failure to pay back a loan, and can be due to a lack of willingness or ability. This might occur with any type of financial debt, bonds, home mortgages, car loans, and promissory notes. When debt obligations are not paid, it can result in great financial hardship for both business and individual. Loan repayment performance is an important factor for a lending institution to measure whether the loan is repaid in full as per the contract terms. Loans with good performance of repayment boost the chances of collecting interest and lower loan losses for lending institutions (Azeez & Ajayi, 2023). However, if loan repayment is poor, it impacts institutional capital, income, and organization goals and can even lead to financial institution failure.

2.2.2 Causes of loan default

Loans to rural borrowers may be associated with structural and livelihood issues that limit their capacity to pay back (Lu et al., 2024). Many households are dependent on rain-fed agriculture and the prices of commodities are not stable in many households, so they are easily affected by weather conditions, crop failures, and seasonal incomes. These factors can cause their repayment plans to be disrupted (Lu et al., 2024; Zhiqiang et al., 2024). Studies conducted among smallholder farmers and sugarcane growers indicate that borrowers' low level of education and financial literacy leads to difficulties in comprehending the terms of their loans, interest rates, and the consequences of default, increasing their likelihood of failing to repay (Kiros, 2020; Lu et al., 2024). Climate risk exposure is also linked to fewer refinancing and restructuring options for borrowers with low or no credit score, few assets to secure loans and high exposure to climate risks (Lu et al., 2024; Zhiqiang et al., 2024). These challenges make rural loans particularly vulnerable to financial shocks and seasonality of income, and this emphasizes the importance of improved risk-sensitive credit assessment (Lu et al., 2024).

2.2.3 The effects of failure to pay loans on financial institutions and rural lenders

The default rates have serious repercussions for the microfinance institutions, rural banks and other lenders for the poor. The converse is true as higher default rate creates more non-performing loans (NPLs) that can cause a drop in interest income and intensify narrow profit margins and capital inadequacy (Adamu, 2020; Afolabi et al., 2020). Defaulting loans in excess of the regulatory limit negatively affect the liquidity, solvency, and obligations-fulfillment of the Nigerian microfinance banks (Adamu, 2020). The more credit risk that is experienced by lenders, the more they have to put into loan recovery, court action, and portfolio monitoring, which increases their operational costs, and leads to funds being diverted from productive lending (Afolabi et al., 2020). High default rates over time make lending and investing in vulnerable rural groups difficult for institutions, forcing them to ration credit, increase the requirements for credit eligibility or simply withdraw from those markets, thereby reducing access for vulnerable borrowers and jeopardizing the sustainability of inclusive finance programmes (Adamu, 2020; Afolabi et al., 2020).

2.2.4 Difficulties of lending in rural areas

Extra challenges arise in rural lending, where it is difficult to measure and monitor credit risk. The informal economy is where many rural borrowers operate, and it is often hard to screen those who wish to borrow and assess their risk before granting the loan, as there may be insufficient documentation, credit history and reliable identification to identify the risk (Igbaria, 1994; International Journal of Economics, Commerce and Management, 2019). On-site appraisal and follow-up visits are more expensive in rural areas and costs are higher as there is less regular supervision in such situations (International Journal of Economics, Commerce and Management, 2019). The financial infrastructure is also weak, such as limited access to digital services, low bank branch density, and limited usage of affordable technology-based risk management tools (An Empirical Analysis of Challenges Faced by Rural Bank, 2023). High environmental and income risk, together with these problems makes traditional lending models unable to price up risk. Consequently, lenders are more conservative, and often denied access to borrowers with potential but no data.

2.2.5 Rural Lending and Financial Inclusion

Smallholder farmers, informal entrepreneurs, and low-income families in rural areas typically do not have regular income, little financial documentation, and are vulnerable to events such as fluctuating prices and unpredictable weather. These factors render it difficult for the traditional lenders to evaluate their credit-worthiness (Balana & Oyeyemi, 2022), thus lending to them becomes difficult. They provide assistance through group lending and social collateral from microfinance institutions and cooperative lending groups. In these systems, borrowers help one another and community members help to monitor, leading to lower default rates and costs. This is because credit is more accessible in areas often overlooked by banks (Umeaduma, 2023; Imran et al., 2022). Nevertheless, there are still many issues to address, including insufficient collateral, financial illiteracy, lack of infrastructure, costly transaction costs, poor credit histories, and the fact that many transactions take place in distant locations. Banks are reluctant to make loans because of these problems (Chen & Xiao, 2025). Increasing financial inclusion is crucial in the context of poverty alleviation, providing support to the rural economy, assisting households to deal with shocks, and boosting the economy through bringing more people into the formal financial sector (Lal, 2019). Due to these problems, there is a need for the development of better credit risk assessment tools. In rural areas with limited information, machine-learning systems using alternative data can assist lenders in making better decisions (Chen & Xiao, 2025).

2.3 Traditional methods for loan default prediction:

Classical prediction methods such as logistic and linear regression, and structured credit scoring methods, were the primary approaches used in traditional loan default prediction. The typical approach for estimating the probability of default for a borrower is logistic regression and its input variables are characteristics of the borrower and the loan. The benefit of this approach is the transparency of the odds ratios, which assists banks and microfinance institutions (MFI) in making decisions. (Costa et al., 2020) Logistic regression remains popular in consumer and retail credit risk for it offers clear results, well-known statistical rules and is easily applied in regulatory or everyday use. Linear regression isn't a good model for a yes or no default prediction, but it has been applied in the past for modeling continuous credit risk measures such as loss given default or delinquency score. It has also played a role in developing early scoring systems that transformed predictions into risk grades or acceptance thresholds (Soomro et al., 2024). Based on these techniques, traditional credit scoring models and score cards used borrower information such as income, job security and repayment history and converted it to points. These scores grouped applicants into risk groups, which facilitated consistent decision making on lending across branches and portfolios, based on policy (Hlongwane et al., 2024; Wu & Pan, 2021). The traditional methods emphasized simplicity, stability, and statistical relationships, fitting the requirements of most supervisors and internal risk managers in financial institutions.

Many banks have traditionally relied on rule-based systems, in which domain experts and credit officers have written lending policies in the form of “if–then” rules to evaluate applications and to manage the risk of the lending portfolio (Hlongwane et al., 2024). These integrated systems were based on expert opinion and were supplemented by regression models or scorecards. For instance, the decision flow could include automatic rejection of applicants whose derogatory marks are sever, and override approved instances in the presence of certain red-flag conditions, built into decision logic, incorporating institutional experience and regulatory requirements (Hlongwane et al., 2024). Traditional methods are interpretable, easy to explain to regulators and customers, and require less computation, but they are limited by their assumptions of linear models and coarse grouping of variables, as well as their inability to capture nonlinear relationships among the variables in large and high-dimensional datasets. While logistic regression and scorecard models are still popular choices, they can sometimes struggle to outperform contemporary machine learning techniques like tree-based ensembles and deep learning when dealing with complex credit risk assessments that factor in a multitude of behavioral, macroeconomic, and transactional data (Soomro et al., 2024). The challenges of dealing with complex patterns, adjusting to changing data and dealing with large portfolios have spurred the adoption of machine learning based loan default prediction systems, especially when there is more data and more computational power (Tanya, 2025).

2.4 Concept of Machine Learning

The term "Machine Learning (ML)" was coined by Arthur Samuel, who worked on the early development of computer learning, a research area that aims to make computers learn from data without being explicitly programmed. Machine Learning is a subcategory of AI which tries to comprehend the structure of the data, and integrate it into models that can be used by humans (Setiawana et al., 2019; Wisdom et al., 2025). Supervised Learning is a learning process in which the algorithm learns a function that transforms an input into an output, using a set of labeled data, and improves the function based on the error. Unlike supervised learning, unsupervised learning aims to classify unlabeled data and to discover structures in data, such as patterns in the input data (Alhasan et al., 2020). The third type is reinforcement learning which concentrates on maximising rewards by training an agent to explore its environment.

2.4.1 Overview of Supervised ML in Loan Default Prediction

Supervised learning for binary classification is frequently employed in loan default prediction. Binary classification assigns new observations to one of two categories and is used to train labelled data in a binary format, such as true/false or positive/negative (Zhi Zheng Kang et al., 2025). Due to their widespread use in binary classification tasks, four supervised learning algorithms, Decision Tree, Random Forest, XGBoost, and LightGBM, are discussed in the following sub-sections.

2.4.2 Decision Tree

A Decision Tree algorithm partitions data into progressively smaller subsets based on attribute values, continuing this process until a specified stopping criterion is met (Nkambule et al., 2024). According to Bansal et al. (2022), the Decision Tree algorithm consists of internal nodes that represent branching structures, datasets that inform algorithmic decisions, and leaf nodes that indicate final outcomes. The algorithm includes two types of nodes: (1) decision nodes, which make decisions and have multiple branches, and (2) leaf nodes, which represent final outputs and do not branch further. Decision Trees can efficiently process large datasets due to their systematic partitioning mechanism, which divides the dataset into smaller segments (Nkambule et al., 2024). However, their applicability is often restricted to straightforward attribute-value data and may require adaptation for more complex scenarios.

2.4.3 Random Forest

Random Forest is a flexible algorithm that can be used for both classification and regression problems. It combines several classifiers to improve model performance and solve complex prediction tasks using ensemble learning. The algorithm creates multiple decision trees and uses them as a meta estimator by applying these trees to different parts of the dataset during training (Sathish et al., 2022). RF works quickly and handles large or imbalanced datasets well. However, it can have trouble training on very diverse datasets, especially for regression tasks (Tay et al., 2024).

2.4.4 XGBoost

XGBoost is a boosting-based tree algorithm known for its optimisations in four areas, which are first a distributed weighted square graph method for segmentation point selection, second, an improved handling of sparse data, third, an efficient cache-aware block data storage structure, and fourth is the better use of parallel and distributed computing (Gao et al., 2021). XGBoost builds trees by splitting leaves within the same layer, which helps reduce overfitting. However, too many leaf nodes can decrease splitting gain and increase computational overhead. In addition, model training requires extensive parameter tuning and ongoing hyperparameter optimisation (Cheng, 2021).

2.4.5 LightGBM

LightGBM is a gradient-boosting decision tree technique that identifies segmentation points and selects features using the histogram algorithm (Liu et al., 2021; Hao et al., 2022). LightGBM uses gradient-based one-sided sampling (GOSS) to prioritise training on samples with low prediction performance and reduce the sample size in each iteration. It applies exclusive feature bundling to lower computational complexity by grouping mutually exclusive features. This method works well because many datasets contain large proportions of empty (or "sparse") data, and some feature attributes cannot have any common values at the same time (Zhi Zheng Kang et al., 2025).

2.5 Review of related works

Odegua (2020) built a prediction model for identifying possible future definers of bank loans using the Extreme Gradient Boosting algorithm. This system allows a bank to quickly identify applicants considered risky when applying for a loan due to past financial history. To build the model, three types of loan-related datasets were used: borrower demographics, borrower repayment history, and borrower historical data provided by Data Science Nigeria through the Zindi platform to train an

Extreme Gradient Boosting classifier to predict defaulting borrowers. In order to test the model's performance, hyperparameter optimization techniques such as grid search and five-fold cross-validation were carried out. In summary, the evaluation of the model resulted in an overall accuracy of 79 percent, a precision of 97 percent, a recall of 79 percent, and a F1 score of 87 percent.

Hota et al. (2025) analyzed bank customer data to determine the most effective machine learning model suitable for predicting loan approval in banks. The study collected a dataset from Kaggle that has attributes describing borrower demographics and loan details to train multiple machine learning algorithms, which include; Naive Bayes (NB), Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), XGBoost, AdaBoost, and LightGBM. Metrics such as Accuracy (ACC), Area Under Curve (AUC), and KS statistics were considered for the model's performance evaluation. Results reveal that the best model is the Random Forest with an accuracy of 88%, an AUC 87%, and a KS 51%.

A review conducted by Alvi et al. (2024) on default prediction models, considering articles published between 2015 and 2024, to identify trends, strengths, and weaknesses, applied a five-step methodology, which consists of scope definition, search strategy, inclusion/exclusion criteria, quality assessment, and result reporting. The research conducted a comparative analysis of statistical; machine learning, deep learning, hybrid, and textual modeling approaches used in loan default predictions. Findings from the study indicated high accuracy, ranging from 80-91% in instances of the hybrid model approach; however, some limitations affect existing systems, such as feature selection challenges, sample size issues, and variability in model performance across datasets.

Muddu et al. (2026) embarked on research that is focused on enhancing credit scoring and loan default prediction in Uganda using integrated financial data and machine learning models to support credit decision-making in rural communities, which are underbanked. The study applied ML techniques, including ensemble methods, to train XGBoost, Logistic Regression, Decision Tree, and Support Vector Machine (SVM) using integrated transactional and loan datasets from fintech companies and telecom providers. Problems of class imbalance in the datasets were addressed through SMOTE oversampling, feature engineering, and integrated data construction was also carried out as part of data preprocessing. Findings from the study, after evaluation, show that the XGBoost showed a remarkable performance accuracy in predicting credit defaulters when compared to other models.

Luo (2025) utilized Machine learning classification techniques with ensemble and boosting algorithms to predict credit default risk and identify key influencing factors. The LendingClub platform dataset was utilized in this experimental study to train and compare multiple ML models, that is, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Classifier (GBDT), XGBoost, AdaBoost, and Bagging, and balancing techniques were applied to retain the best model and improve performance. The result from the study illustrates Gradient Boosting Classifier as the best model with an AUC of 0.691, an accuracy of 0.84, and a recall for the default class initially 0.04 but improved to 0.410 after Random Over Sampler balancing.

Enhancing credit risk in supply chain finance by Zhou and Wang (2025) through applying interpretable machine learning models and SHAP for explainability. The study carried out a comparative evaluation of multiple ML models for Supply Chain Finance (SCF) credit to determine the best model to adopt and integrate the SHAP explainable AI. The model performance was evaluated considering accuracy, precision, recall, F1-score, and the XGBoost had the best performance achieveing an average Precision 92.8%, Accuracy 91.9%, Recall 94.7%, and F1-score 93.7%, outperforming RF, LSSVM, and CNN.

Ozkurt (2024) improved the prediction of personal loan eligibility using ML techniques so as to support financial institutions in making better lending decisions. The study trained and evaluated several boosting-based ML models, such as Gradient Boosting, XGBoost, AdaBoost, and LightGBM on a customer banking dataset to determine the best model for loan eligibility classification. The results from the model evaluation, considering accuracy, precision, recall, and F1-score indicates that the XGBoost outperformed others with an accuracy of 0.95, precision of 0.95, recall of 0.95, and F1 of 0.95.

Akinjole et al. (2024) developed an ensemble-based machine learning framework for predicting credit default risk and identifying optimal preprocessing and modelling techniques. The methodology applied in this work involves a multi-stage pipeline that includes feature selection, data cleaning, class imbalance handling, training of the model, and ensemble learning. The dataset for this study was collected from a public dataset in the Kaggle online repository called the LendingClub. Subsequently, after extensive data preprocessing and hyperparameter tuning, it was used to train the Random Forest, Decision Tree, SVM, XGBoost, AdaBoost, and Multi-Layer Perceptron (MLP) for accurate credit default risk prediction in an ensemble approach that involves voting and stacking. Findings from the study indicate that the best ensemble approach was the stacking, which had an accuracy of 93.69%, precision of 95.59%, recall of 95.55%, and AUC of 97.81–98%.

Monje et al. (2025) developed an ML model with explainable AI (XAI) capable of predicting early loan defaults while providing interpretable explanations for decision-making. The study utilised a real-world large P2P loan dataset from LendingClub to train multiple ML algorithms consisting of

Logistic Regression, Decision Tree, Random Forest, Balanced Bagging, XGBoost, KNN, Gaussian Naïve Bayes, Neural Networks, and AdaBoost for early loan default risk prediction. XGBoost achieved an AUC of 0.731 and an accuracy of 0.795, outperforming other algorithms.

Puspitasari et al. (2025) collected secondary data from audited financial statements and health rating assessments from the Indonesian Financial Services Authority to develop a predictive framework for assessing the sustainability of rural banks in Indonesia in the post-COVID-19 period using three logistic regression models. The study applied a statistical modelling technique that considered Stepwise logistic regression with AIC minimisation, multicollinearity testing using the Variance Inflation Factor (VIF), cross-validation, and robustness checks. A Comparison across three logistic regression model specifications, Model 1 (CAR, ROA, BOPO, LDR), Model 2 (NPL, CR, OBS), and Model 3 (combined variables), was conducted and the findings indicate the Model 3 demonstrated the strongest predictive capability with higher logit scores, for instance the validation average logit of 0.9690 exceeding cut-off 0.9605, indicating that it has a strong classification accuracy.

Recently, Zhang et al. (2025) investigated the application of several machine learning techniques to predict loan default on a real-world lending dataset. The main goal was to evaluate the ability of these models to predict and to benchmark their performance against traditional loan assessment techniques. For this, researchers conducted an extensive machine learning workflow, which included data preprocessing, feature engineering, class imbalance handling, model development, hyperparameter optimization, and model evaluation. The four ensemble learning algorithms, Gradient Boosting, XGBoost, Random Forest and LightGBM were trained and compared under the identical experimental conditions. The evaluation showed that the Gradient Boosting model exhibited the highest overall predictive performance with an accuracy of 88.87%, an F1 score of

80.84% and a recall score of 80.21%. However, XGBoost showed the best discriminative ability with a ROC-AUC score of 97.14%, which means it had the best performance in differentiating between defaulting and non-defaulting borrowers.

Likewise, Hossain et al. (2025) suggested a secure machine learning model for bank loan prediction for a public available Kaggle loan prediction dataset. The goal of the study was to design an automated loan prediction system that ensures data privacy by combining differential privacy with machine learning. The experimental framework was modified by introducing two privacy-preserving techniques: the Laplacian and the Gaussian methods, and evaluated with five classification algorithms: Random Forest, XGBoost, AdaBoost, Logistic Regression and CatBoost. In the experiments, a 10-fold cross-validation method was used and the differential privacy budget was systematically varied to understand its impact on predictive performance. The results showed that the Random Forest classifier showed the highest accuracy of 62.31% for privacy budget 2 in the Laplacian privacy mechanism. In contrast, with the privacy budget of 1.5, the CatBoost model demonstrated the best performance among all other models achieving accuracy of 81.25%, which showed its capability for a privacy-conscious loan prediction task.

Table 2.1 Summary of Related Literature

Author and Year	**Technique**	**Work Done**	**Limitations**
Odegua (2020)	Machine Learning (ML)	developed a predictive model that identifies bank loan defaulters using the extreme gradient boosting algorithm, helping the institution pinpoint risky loan applicants and reduce financial losses, and the results showed an accuracy of	The system was not deployed in real time, and its accuracy can still be improved.

		79%, precision of 97%, recall of 79%, and an F1-score of 87%.	
Hota et al. (2025).	Machine Learning	The study collected a dataset from Kaggle that has attributes describing borrower demographics and loan details to train multiple machine learning algorithms, which include; Naive Bayes (NB), Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (F), XGBoost, AdaBoost, and LightGBM. Random Forest had best result with an accuracy of 88%, an AUC 87%, and a KS 51%.	The system was not deployed in real-time.
Alvi et al. (2024).	Literature review	The research conducted a comparative analysis of statistical, machine learning, deep learning, hybrid, and textual modelling approaches for loan default prediction, with findings indicating high accuracy for the hybrid model, sample size issues, and variability in model performance across datasets.	Survey study and no experimental validation

Muddu et al. (2026)	ML; Ensemble methods	The study conducted research focused on enhancing credit scoring and loan default prediction in Uganda, using integrated financial data and machine learning models to support credit decision-making in underbanked rural communities, and the XGBoost showed a remarkable performance accuracy in predicting credit defaulters when compared to other models.	Limited dataset from a few financial operators and no online/real-time deployment
Luo (2025)	ML; Ensemble methods	The work utilised machine-learning classification techniques, including ensemble and boosting algorithms, to predict credit default risk and identify key influencing factor, with Gradient Boosting Classifier performed best with an AUC of 0.691, an accuracy of 0.84.	The dataset has limited features and no real-time deployment
Zhou and Wang (2025).	ML; XAI	The study conducted a comparative evaluation of multiple ML models for Supply Chain Finance (SCF) credit to determine the best model to adopt and integrated SHAP explainable AI. The	Lacks real-time deployment and a limited sample size

		evaluation result indicates that XGBoost had the best performance	
Ozkurt (2024)	Machine Learning	The study trained and evaluated several boosting-based ML models on a customer banking dataset to determine the best model for loan eligibility classification. The results indicate that XGBoost outperformed the others, achieving an accuracy of 0.95, precision of 0.95, recall of 0.95, and F1 of 0.95.	Offline predictive analysis for decision support in loan and no real-time deployment
Akinjole et al. (2024)	Ensemble approach; ML	The developed an ensemble-based machine learning framework for predicting credit default risk and identifying optimal preprocessing and modelling techniques. Findings from the study indicate that the best ensemble approach was stacking, which achieved an accuracy of 93.69%, precision of 95.59%, recall of 95.55%, and an AUC of 97.81–98%.	model optimization complexity and no real-time implementation
Monje et al. (2025)	ML; XAI	The study utilised a large, real-world P2P loan dataset from LendingClub to train multiple ML algorithms with explainable	Black-box nature of advanced models requires explainability methods

		AI (XAI) capable of predicting early loan defaults and providing interpretable explanations for decision-making. XGBoost achieved an AUC of 0.731 and an accuracy of 0.795, outperforming other algorithms.	
Puspitasari et al. (2025)	Statistical Approach	The study applied a statistical modelling technique that considered Stepwise logistic regression with AIC minimisation, multicollinearity testing using the Variance Inflation Factor (VIF), cross-validation, and robustness checks.	The dataset cannot be publicly shared due to institutional confidentiality and no real-time deployment
Zhang et al. (2025)	ML	The study trained multiple ML models on a real-world loan dataset to evaluate their effectiveness in predicting loan defaults and to determine how they might outperform traditional approaches. Gradient Boosting had highest overall classification performance in terms of accuracy and F1-score	Class imbalance in the loan datasets and no real-time implementation
Hossain et al. (2025).	Differential privacy;	The study trained machine learning algorithms for bank loan prediction using	Data scalability issues.

	Machine Learning	a publicly available Kaggle dataset. The analysis shows that when the Laplacian mechanism with a DP budget of 2 is used, the RF model achieves the highest accuracy of 62.31%.	

2.6 Research Gap

The study has reviewed many existing studies on loan default risk prediction systems, and some of these studies have developed and proposed AI-based solutions that help financial institutions mitigate the risk of credit defaulters. However, most of these studies present experimental results with no real-time deployment or validation, and issues related to data availability and quality are observed as persistent challenges in most of the reviewed studies. Therefore, the study proposes the design of a robust Machine Learning-based solution that predicts loan default risk, along with web-based software integration for real-time validation.

CHAPTER THREE

SYSTEM ANALYSIS AND DESIGN

3.0 Introduction

This chapter describes the system analysis and design of the proposed Machine Learning-Based Loan Default Prediction for Individuals in Rural Cooperative Societies. It reviews the existing manual and intelligent system used in loan default prediction and identifies challenges affecting them. The chapter also presents the methodology, system requirements, and design implementation for developing the proposed system.

3.1 Analysis of the Existing System

The existing system involves a manual loan evaluation method, a basic rule-based credit scoring system and traditional banking credit systems. In this method, a person looking for a loan contacts either the banking officials directly or by using a cooperative group. The person completes an application that shows personal info, income level & then gives the bank info on how much they would like to borrow, what will this new loan will be used for. The person will then provide the loan officer with proof that they are who they say they are- commonly called ID verification; i.e. by providing their national ID number, proof of their income, a receipt showing membership in the cooperative group. This verification process allows the loan officer to determine whether or not the applicant is a valid loan applicant according to the credit criteria. As it pertains to the current process, the loan officer uses human judgement to evaluate the overall creditworthiness of the person applying for a loan. To do this, the officer will review various, manually determined factors: the borrower's ability financially to repay this loan, prior history of repayment on past loans, & the number of cooperatives that member borrowing through their cooperative has repaid in the past. This is the process the loan officer carries out manually in order to determine whether the individual qualifies for the loan, then, before proceeding into collateral or guarantee check, but for the scope of this research, the cooperative society serves as the guarantor. In the same vein, current rule-based systems utilise the same process that is programmed using if-else statements. However, these rule-based systems are not intelligent and cannot learn from historical data.

3.1.1 Problems with the existing system

1. The existing manual systems are biased and inconsistent, as different officers may make different decisions regarding a particular applicant, and personal judgment can also affect outcomes.

2. The time it takes to cover multiple loan applicants in a manual approach is slow, therefore delaying loan approval.

3. The existing manual and rule-based system has no predictive intelligence, resulting in it making poor predictions, as there are no data-driven models to enable it to detect hidden risk patterns.

4. They make use of only basic information, making it unreliable, and existing systems ignore complex relationships in data.

5. There is a high default risk in the current manual and rule-based approach, as decisions are not optimised, leading to big financial losses.

3.2 Analysis of the Proposed System

The proposed Machine Learning-Based Loan Default Prediction for Individuals in Rural Cooperative Societies was designed to solve the problems with the existing systems, such as the fact that the existing system is not user-friendly, relies heavily on human judgment, is time-consuming and has no predictive intelligence. It leverages historical financial data to provide accurate, fast and data-driven credit risk assessments. The objective of this proposed system is to predict loan default risk accurately and support rural banks in decision-making, thereby reducing financial losses and improving loan accessibility for low-risk individuals.

The proposed system will employ Extreme Gradient Boosting (XGBoost), a powerful ML algorithm used for classification and prediction tasks on mostly structured data, such as the loan default datasets. XGBoost works through a boosting process of building multiple decision trees sequentially; each new tree corrects the likely errors made by previous trees, and the model improves step by step until it attains an accurate prediction. The new system is designed to receive the borrower's information, including income, loan amount, credit history, employment status, and number of dependents. The XGBoost has been trained on past data using the same parameters and will now classify the loan applicant as Default (1) or No default (0), and assign a probability score indicating the risk of default. XGBoost minimises error using a loss function in addition to regularisation: the loss measures prediction error, while the regularisation prevents overfitting, all in a bid to make the model accurate and stable. The new model will now be integrated into a user-friendly software application via the model's API to enable an easy system flow, where the user just needs to input the loan data, which is then preprocessed, and the XGBoost model processes the input to output a classification result.

3.2.1 Advantages of the Proposed System.

1. The new system eliminates the manual decision-making approach of the existing system by creating an automated approach for decision-making.

2. There is an improvement in prediction accuracy as the intelligent system overrides the manual approach, eliminating factors like bias.

3. Loan processing time is now faster and seamless.

4. The new system is data-driven and solves the problem of different loan officers having different opinions on the same applicant.

5. The new system will be scalable and adaptable, meaning there is room for growth and updates.

3.3 System Requirements

3.3.1 Functional Requirements

i. Users should be able to perform their registration and login.
ii. The system will accept the borrower's financial details.
iii. The system will process input data for prediction.
iv. The system uses XGBoost to predict loan default status.
v. The system should display prediction results and as well store the loan records.
vi. The system should allow the admin to view reports.

3.3.2 Nonfunctional Requirements

i. The system should be able to give reliable prediction results always with a high degree of accuracy with minimum classification errors.

ii. The application needs to be efficient in handling user requests and to provide prediction results with minimum response time for real-time interaction.

iii. Security mechanisms should be included to protect the system from access by unauthorized users, and to ensure that information belong to users is kept confidential and complete.

iv. The architecture should support scalability, enabling the application to handle high volumes of requests from multiple users and grow as needed without sacrificing performance.

v. All user-related information shall be kept in a secure manner through industry standard encryption to avoid disclosure or modification of the data to another party.

vi. A user interface should be responsive and flexible, delivering a consistent and smooth user experience on desktop, tablet, and mobile devices.

3.4 The Proposed XGBoost Algorithm

Machine Learning is used on structured data to make accurate predictions at a high level of precision and is used for tasks like classification primarily. XGBoost is a specific type of machine learning that has been proven to outperform other machine learning algorithms for prediction-related purposes. In this research, the static default predictions that were made using XGBoost were produced utilizing various features from within the structured data, such as income, loan amount, credit history, employment status, and number of dependents.

How XGBoost Works

Step 1: Generate the Initial Prediction

Like most other machine learning algorithms, the XGBoost algorithm starts by predicting a base value for each borrower in the training set. Typically, this initial guess is a result of the overall rate of default that is observed in the dataset prior to building any decision trees in binary loan default classification.

Step 2: Determine the Prediction Errors

The algorithm then measures the difference between the actual outcomes of the loans and the predicted probabilities. These prediction errors (also known as residuals) indicate the amount of data that would need to be changed to make the current model a better predictor and expose observations that need to be addressed.

Step 3: Build a Decision Tree to Correct Errors

A new decision tree is built on the residuals calculated, instead of the original target values. This tree's goal is to identify the trends of past errors in prediction, focusing particularly on borrowers who were misclassified, such as those who were predicted as high-risk but were actually low-risk or those who were predicted as low-risk but were actually high-risk.

Step 4: Update the Model Using the Learning Rate

Now that the learning rate has been determined, it is time to update the model.Once the learning rate has been decided, it is time to update the model. The newly built tree is used to make predictions, which are scaled by a predetermined learning rate and added to the model. A learning rate makes the model learn slowly, improves its ability to extrapolate to new data, and prevents overfitting.

Step 5: Perform Successive Boosting Iterations

Repeat the steps from step 4, but incrementally add one iteration at a time. This process of calculating residuals, training correction trees and updating predictions is repeated in an iterative manner. This study is run for 300 boosting iterations where the successive trees are designed to improve the performance of the previous trees.

Step 6: Produce the Final Classification

Once all the boosting rounds have been completed, the predictions from each of the decision trees are aggregated to create the final prediction. This model is used to generate the estimated probability of loan default and classification label: 1 indicates the borrower is expected to default, while 0 indicates that the borrower is expected to repay the loan successfully..

Input

Training Dataset (Dtrain)

The training dataset is used to construct and optimize the XGBoost model.

Feature Matrix (Xtrain)

The predictor variables include:

- Encoded categorical variables such as gender, marital status, and employment status.
- Financial information, including applicant income and requested loan amount.
- Credit-related attributes, particularly previous credit history.
- Household and demographic characteristics, including education level and number of dependents.

Target Variable (ytrain)

The output variable represents the loan repayment status:

- **1** – Loan Default
- **0** – No Default

Testing Dataset (Dtest)

The testing dataset is used to evaluate the predictive performance of the trained model.

Feature Matrix (Xtest)

The same preprocessing procedures applied to the training data are performed on the testing features to ensure consistency.

Target Variable (ytest)

Contains the actual loan repayment outcomes used to assess model accuracy and other evaluation metrics.

Model Hyperparameters

The XGBoost classifier is configured using the following hyperparameter settings:

- **Number of estimators (n_estimators):** 300 boosting trees
- **Learning rate (learning_rate):** 0.05
- **Maximum tree depth (max_depth):** 6–7
- **Subsample ratio (subsample):** 0.80

- **Column sampling per tree (colsample_bytree):** 0.80
- **Objective function (objective):** Binary logistic classification (binary:logistic)
- **Evaluation metric (eval_metric):** Logarithmic Loss (LogLoss) or Area Under the ROC Curve (AUC)
- **Random seed (random_state):** 42

Output

After training and evaluation, the XGBoost model produces the following outputs:

1. **Predicted Loan Status**
   - Default (1)
   - No Default (0)
2. **Probability Score**
   - A probability value ranging from 0 to 1 that represents the estimated likelihood of loan default. For example, a prediction score of **0.78** indicates a 78% estimated risk that the borrower will default on the loan.

Table 3.1: Systems Input-Process-Output Analysis

This section explains the data flow through the proposed system, presented in Table 3.1.

Stage	**Component**	**Description**
Input Stage	Borrower Financial Data	Includes income level and loan amount requested by the individual
	Credit Information	Historical credit behaviour indicating repayment reliability
	Socio-Economic Data	Employment status and number of dependents of the borrower

	User Interface Input	Borrower data entered through the system interface (web or desktop application)
Processing Stage	Data Preprocessing	Cleaning data, handling missing values, encoding categorical variables, and normalisation
	Feature Selection	Identifying the most relevant features influencing loan default prediction
	Model Training	Training the XGBoost model using historical loan data
	Prediction Engine	The model analyses input data and predicts the likelihood of loan default
Output Stage	Prediction Result	Classification of borrower as Default or No Default
	Risk Score	Probability score indicating the likelihood of loan default
	Result Display	Prediction output presented to the user via the system interface

3.4.2 System Architecture

This system's architecture is created with great care to allow for smooth transfers of data throughout the entire loan default prediction process, from the step where borrowers enter their information (user input) through the step of determining if the borrower has defaulted on a loan. A diagram of the entire loan default prediction system and how data flows through the system is shown in Figure 3.1.

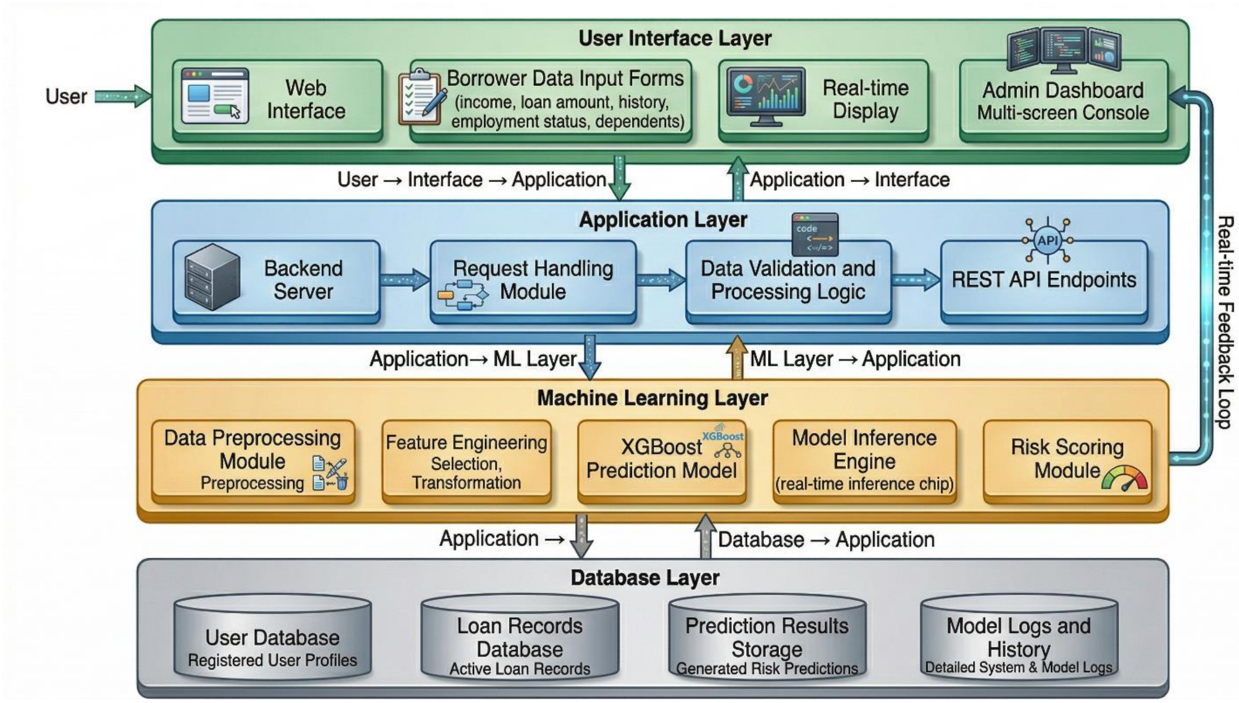

Fig 3.1 Architectural diagram of the loan default prediction system

The user interface, which is the first place people will use to input their financial information (or where they will see the results of the loan default prediction) is separate from the application layer, which ensures that all data entered has been securely validated and that data will have been moved between application components in a trusted manner. The application's machine learning layer processes and transforms the data received from the application layer prior to using the XGBoost algorithm to make an accurate risk prediction for each borrower. Once the prediction has been received, the application will forward it back to the user via the user interface. Finally, supporting the entire system is a database registration layer that contains all user data, loan history, and prediction results; this information is stored securely, allowing for the protection of user data and allowing for the analysis of long-term trends in the data so that informed decisions can be made by cooperative societies about the borrowers they serve.

3.5 SYSTEM DESIGN APPROACH

3.5.1 Proposed Methodology: Object-Oriented Analysis and Design

The proposed machine learning based loan default prediction system has been developed using methodology of Object-Oriented Analysis and Design (OOAD). The approach is to create a system of interacting objects that model real-world entities as a basis for designing a software application: a systematic approach. It's modular architecture, reusability, maintainability, and scalability make it an ideal fit for intelligent software systems that integrate predictive analytics with contemporary application development. In order to make the system architecture user oriented and structured, the system design is done using Unified Modelling Language (UML), and system implementation is done based on Object Oriented Programming (OOP). All of these methods enable accurate system documentation, software development and future improvements.

The first step in the methodology is the definition of the requirements: functional and non-functional requirements are gathered in this phase of the methodology. These requirements come from an evaluation of the constraints of the available methods of assessing loans, coupled with the needs of financial institutions, loan officers, loan administrators and loan applicants. The results of this phase will be used to establish the functions that should be provided by the intelligent prediction platform.

Next, the modeling of the overall design and interaction between the system's components (using UML diagrams) is performed: system design. Each illustrates a different view of the proposed application and includes use case, class, activity, and sequence diagrams. The overall purpose of these diagrams is to represent the relationships among the actors, the prediction engine, the database, and other components, as well as the static structure and dynamic behaviour of the system.

The next phase of the design is the system implementation in which the conceptual models are converted into executable software by implementing them in an object-oriented manner. The predictive engine is developed in Python using the Flask framework, which incorporates the trained XGBoost model for loan default classification. The client-side interface is designed with React Native to create a fast and intuitive application that will be accessible on different devices.

The methodology ends with system testing, deployment and maintenance. The app's functionality, prediction, security, and usability are tested during testing to make sure that it meets the specified requirements. After this it is deployed for actual operation and regular maintenance procedures are used to enhance its performance, resolve software problems and to incorporate future changes.

3.5.2 Unified Modelling Language (UML) diagrams

Unified Modelling Language (UML) diagrams are used to represent the proposed system in a visual way with respect to its architecture, functionality & operation. They provide an essential tool for design, showing how various software elements work together in order to enhance communication between software developers, researchers and other stakeholders during software development. In the case of the proposed machine learning based loan default prediction system, there are a number of UML models created that describe the structural and behavioural characteristics of the system. These models all show the interactions between the users, the prediction module, the database, and other components of the supporting system, and offer a blueprint that can be followed when implementing.

A. Use Case Diagram for the Proposed System

The Use case diagram provides an overview of the functions of the proposed system from the users point of view by identifying the major actors and the services provided for each of them. It gives an overview of the interactions between external users and the application in order to achieve certain goals. As shown in the above-mentioned Figure 3.2, the key roles in the proposed loan default prediction system are of Borrower, Loan Officer, and, Administrator. Every actor has their own specific duty in the application, such as entering loan details, requesting loan risk evaluation, viewing predicted results, operating user accounts and operating system resources. The use case diagram thus sets the boundaries of the functional scope of the application and specifies the interactions needed to facilitate intelligent loan default prediction.

.

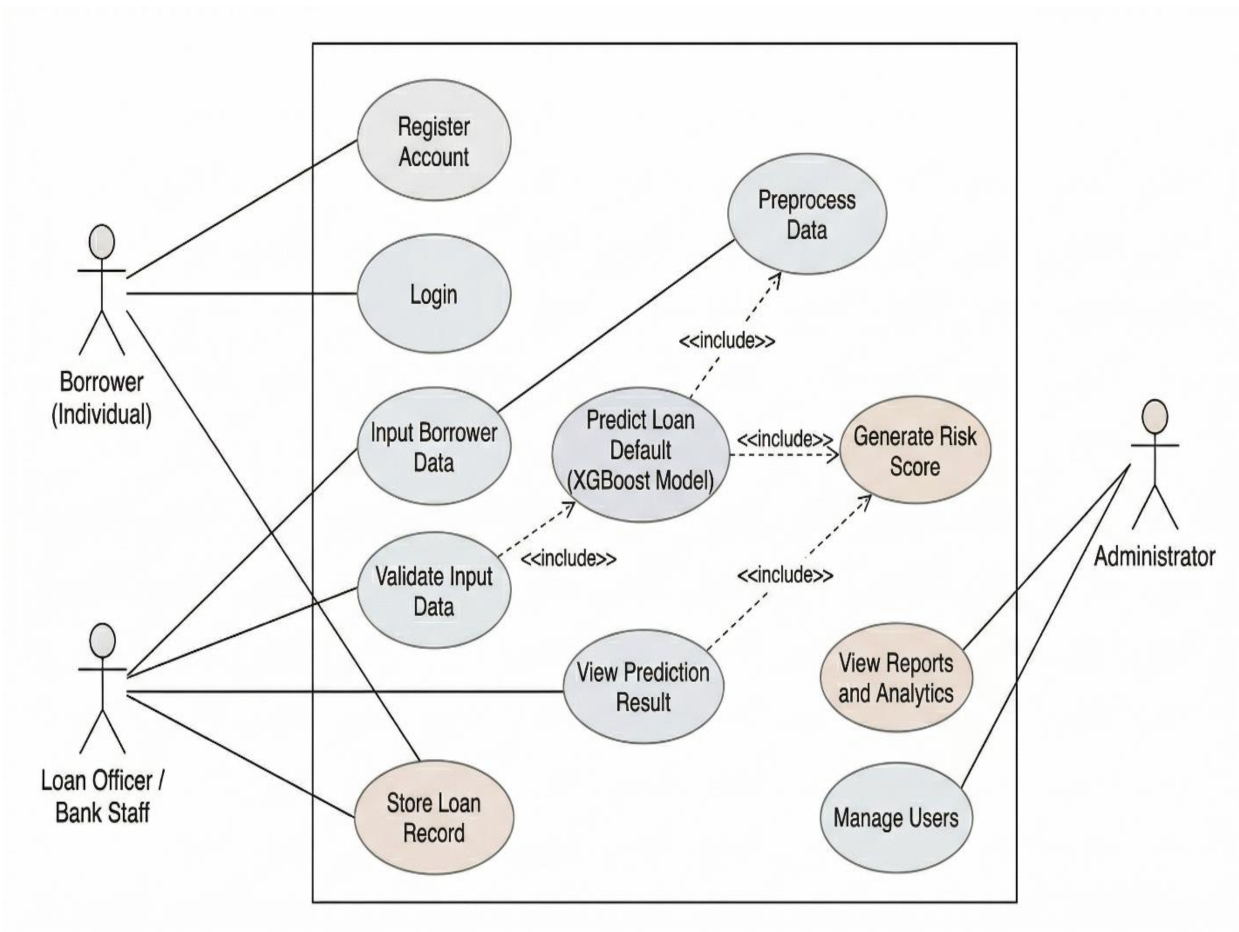

Figure 3.2 Use case diagram of the ML-based Loan Default Prediction

Figure 3.2 shows the proposed system's use case, where the borrower can register their personal details, gain secure access to their profile, and view results. The Loan officer representing administrative users from the lending bank has the primary function of data management, which involves entering applicant data, validating the entered data for accuracy and permanently saving the application (storing the loan record). The administrator handles high-level system maintenance and possible system oversight.

B. Class Diagram

The UML class diagram illustrates the system's static structure, showing the blueprint of the objects, their data, and how they interact. Figure 3.3 presents the class diagram of the machine-learning-based loan default prediction system for individuals in a rural cooperative society. It is made up of seven (7) classes.

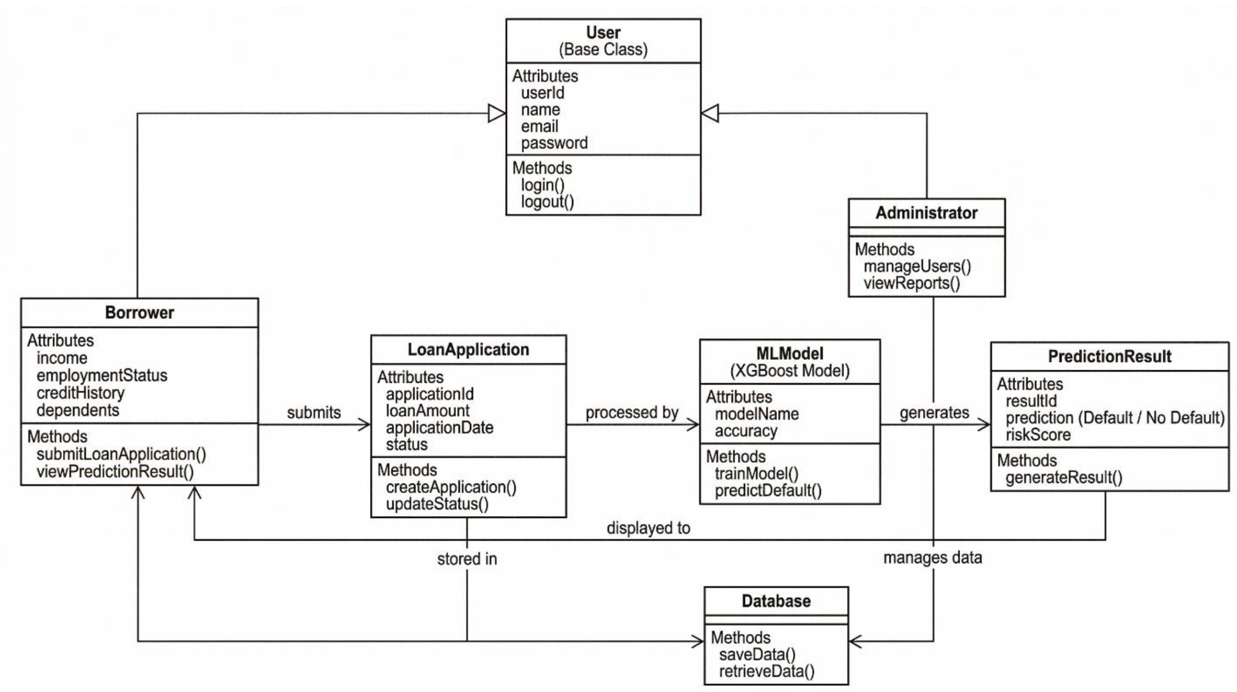

Figure 3.3: Class Diagram of the Loan Default Prediction System

The User is the base class containing common credentials like userId and email, along with basic login and logout functionality, while the borrower class inherits from the User and stores specific socio-economic data, such as income and dependents, and also enables loan submission. The Administrator class likewise inherits from the user class, but it focuses on system oversight, user account management, audit reporting, and system security. The other classes are LoanApplication, MLModel, PredictionResult, and Database.

C. Activity Diagram of ML-based Loan Default Prediction System

The activity diagram is just like a simple flow diagram, which shows the flow, activities, and decisions in a step-by-step manner within a particular process in the Loan Default Predictor. Figure 3.4 shows how the user signs up or logs in to the system and can view whether they are eligible for a loan based on the classification results.

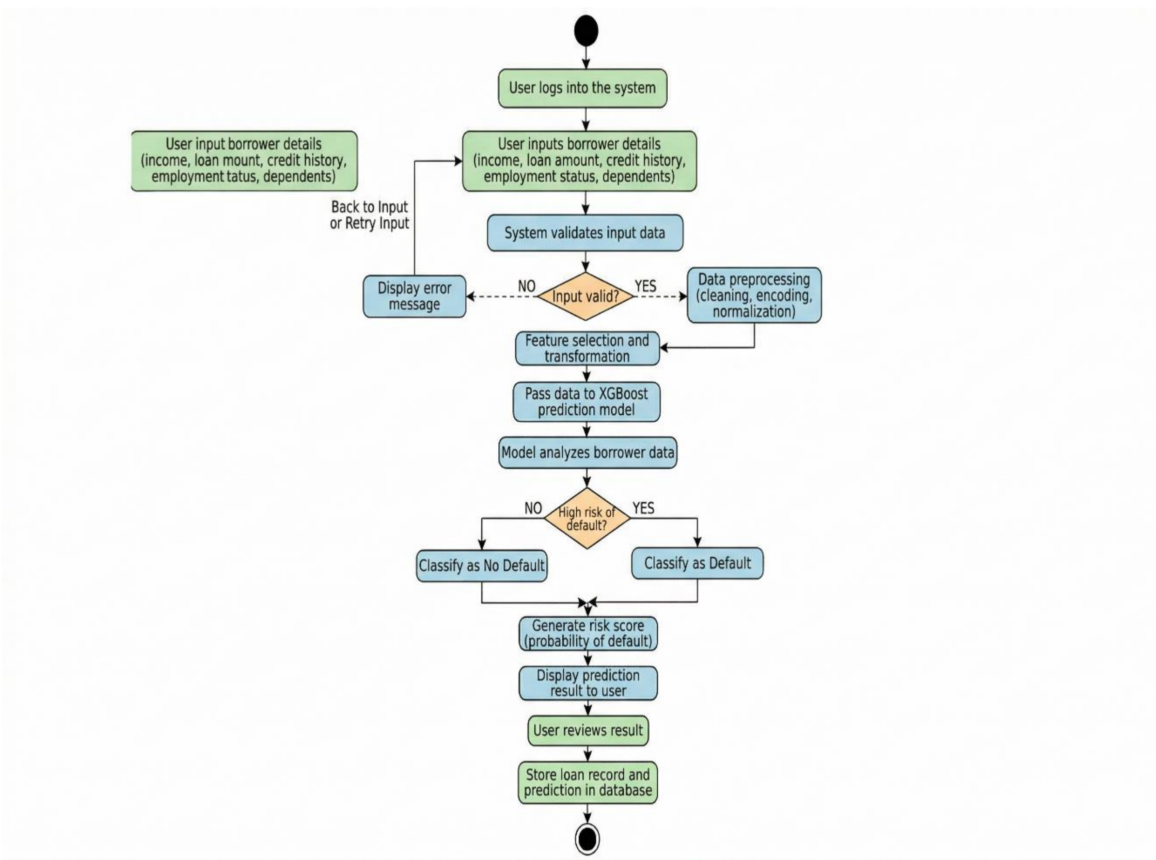

Figure 3.4 Activity diagram

According to Figure 3.4, the process begins with user authentication and the entry of borrower details, after which the system verifies that the input is complete and accurate, redirecting the user to make corrections if necessary. Once validated, the data is cleaned and transformed through preprocessing and feature selection before being passed to the XGBoost model for analysis and evaluation of the likelihood of default, and, based on a defined risk threshold, the model classifies the borrower as either "Default" or "No Default." A probability-based risk score will be generated in order to be displayed to the user, who will be tasked with reviewing the result before it is stored in the database, which marks successful completion of the transaction.

3.6 Data Collection and Preprocessing

To create a loan default prediction system, multiple sources of data are incorporated for improved accuracy and reliability of the model. The loan default datasets were sourced from various locations online including Kaggle where a combination of demographic information (e.g., age) and monetary information (e.g., amount of the loan, income) could support the objective of predicting loan defaults. The original quality and size of the datasets was improved substantially during preprocessing. The original dataset had 307,511 records and 122 fields, complied from Home Credit datasets. After columns with more than 50% of their records missing were removed, the number of fields associated with the dataset fell to 80 while maintaining all 307,511 records. This removal of noisy and untrustworthy fields from the dataset reduces any negative impacts they may have had on model performance and increased computation time for modeling purposes. Further refinement of the datasets included selecting the most relevant fields to the rural cooperative societies' ability to repay loans by retaining only those fields that focused on income level, loan amount, repayment ability, type of occupation and size of household, and renaming variables to make them clearer and to better relate to the rural cooperative societies (e.g., changing

AMT_INCOME_TOTAL to "Income" and AMT_CREDIT to "Loan_Amount"). The subsequent improvements made to the dataset provided substantial quantity, quality, and structure improvements to the data, which qualified the dataset for modeling.

3.7 Database Design

A. User Database

Table 3.2: User Table for Loan Default Prediction System in Rural Cooperative Societies

Attribute	Data Type	Description
id	Int	Unique identifier for each user
name	Varchar	Full name of the user (borrower, loan officer, or administrator)
username	Varchar	Username chosen by the user
email	Varchar	User’s email address
password	Varchar	Hashed password for authentication
role	Enum	Specifies user role (Borrower, Loan Officer, Administrator)
cooperative_name	Varchar	Name of the cooperative society the user belongs to
address	Varchar	User’s residential location (rural area)
is_verified	Boolean	Indicates whether the user account has been verified
date_joined	Timestamp	Timestamp when the account was created
updated_at	Timestamp	Timestamp of last update to user information
reset_password_otp	Varchar	One-time password used for password reset

reset_password_expiry	Timestamp	Expiration time for password reset OTP

B. Loan Prediction Database

Table 3.3: Loan Prediction Table for Individuals in Rural Cooperative Societies

Attribute	Data Type	Description
id	Int	Unique identifier for each loan prediction record
user_id	Int	Foreign key linking to the borrower in the User table
cooperative_name	Varchar	Name of the cooperative society associated with user
income	Float	Individual borrower's income level
loan_amount	Float	Amount of loan requested by the borrower
credit_history	Int	Credit history indicator (1 = good, 0 = poor)
employment_status	Varchar	Employment status of the borrower
dependents	Int	Number of dependents
prediction_result	Varchar	Model output classification (Default / No Default)
risk_score	Float	Probability score indicating likelihood of default
model_used	Varchar	Machine learning model used (e.g., XGBoost)
remarks	Varchar	Explanation or reasoning behind the prediction
created_at	Timestamp	Timestamp when the prediction was generated

Table 3.4 Cooperative Society Table

Attribute	Data Type	Description
cooperative_id	Int	Unique identifier for each cooperative society
cooperative_name	Varchar	Name of the cooperative society

location	Varchar	Rural location of the cooperative
total_members	Int	Number of registered members
created_at	Timestamp	Date the cooperative was registered in the system

CHAPTER FOUR

SYSTEM IMPLEMENTATION

4.1 Introduction

There are multiple sections in this chapter that provide a description of the implementation of the machine learning system for predicting loans default. These sections will detail both the software tools used to build it and how the model was created and evaluated.

4.2 Choice of Programming Languages

4.2.1 Backend

The proposed machine learning application is being built with the Python Flask programming framework to create a back-end and support for RESTful web applications. Flask was selected for its lightweight nature with limited functionality, easy integration with ML libraries that are also written with Python, and also because it provides routing of the application, handles client requests, and manages communication with the trained XGBoost prediction model.

4.2.1 Frontend

React Native

The application will utilize React Native to provide users with an attractive user interface while providing a cross-platform solution using a single source code base. The choice of React Native provides the ability to create a user interface that looks similar and operates like a Mobile Application on Android or iOS, as well as reusability of the components that make up the user interface, which will reduce the time necessary to develop the application and improve maintenance.

4.2.3 Database

MongoDB

The application will be built with MongoDB as the database management system. As MongoDB is a NoSQL database, it provides the greatest flexibility in storing user profile details, loan application datasets, and prediction results in easily stored 'flexible' documents (BSON). The MongoDB database will be developed in the cloud using the MongoDB Atlas database platform.

4.3 Choice for Model Development

For creating the predictive model, the XGBoost classifier was utilised because of its outstanding ability to complete structured classification tasks. The prepared dataset was split into training and testing data sets; both the training and test sets were used in developing the predictive model, which would subsequently be evaluated through hyperparameter tuning in order to enhance the performance of the predicted outputs and to reduce overfitting. The predictive performance of the model was then evaluated using standard classification metrics such as accuracy, confusion matrix and Receiver Operating Characteristic (ROC) curves; the final model achieved a total of 85.4% of accuracy with respect to prediction accuracy and therefore demonstrated a high level of accuracy for identifying borrowers that potentially defaulted on their loans, despite the fact that the dataset was imbalanced with respect to class. The accuracy of the validation of the trained model is shown in Figure 4.1.

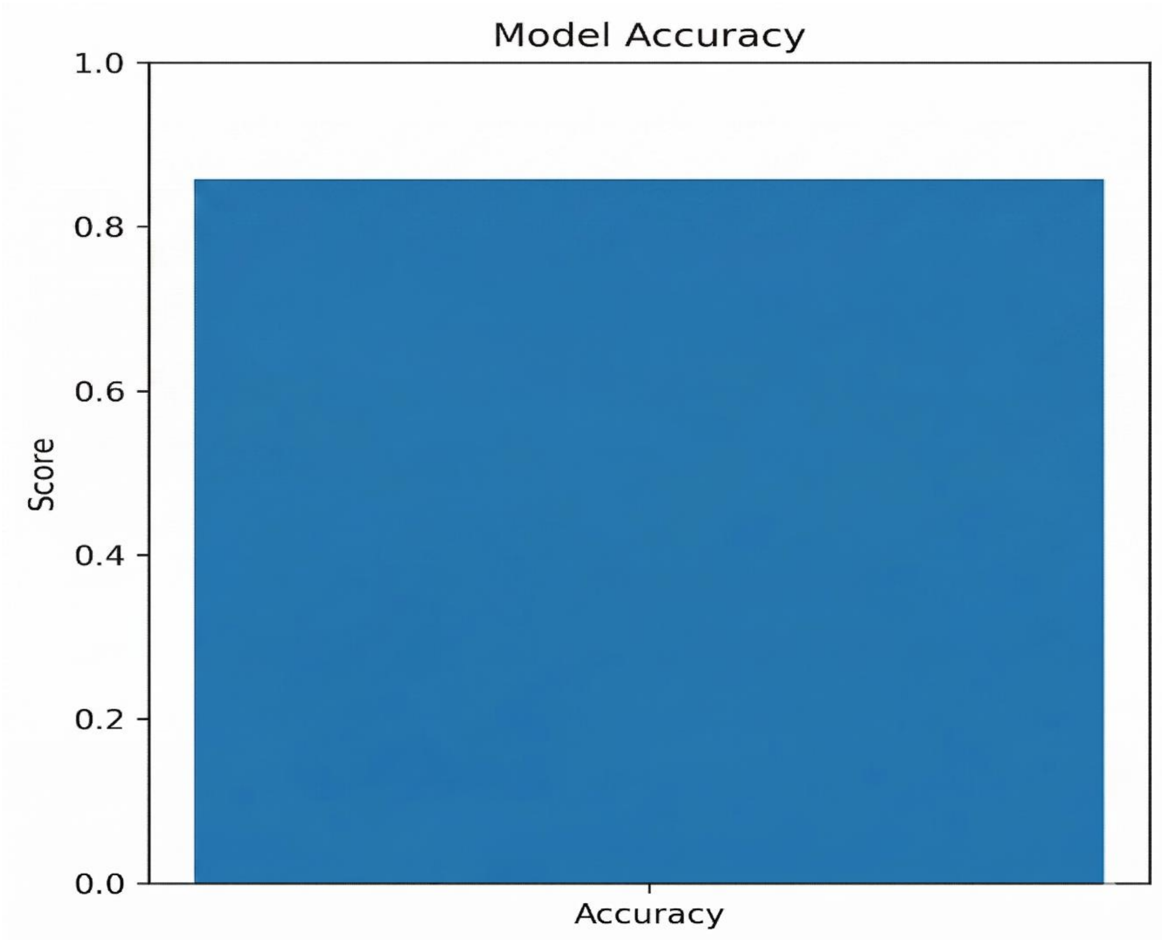

Figure 4.1 Model accuracy

4.3.1 Confusion Matrix Analysis

The trained XGBoost classification model's classification performance was assessed by comparing predicted results to the actual loan statuses through the creation of a confusion matrix. It can be used to provide a detailed summary of how many instances were correctly classified as well as incorrectly classified, thereby providing insight into the model's ability to differentiate between borrowers that defaulted and those that did not default; The confusion matrix created using the trained model is shown in Figure 4.2.

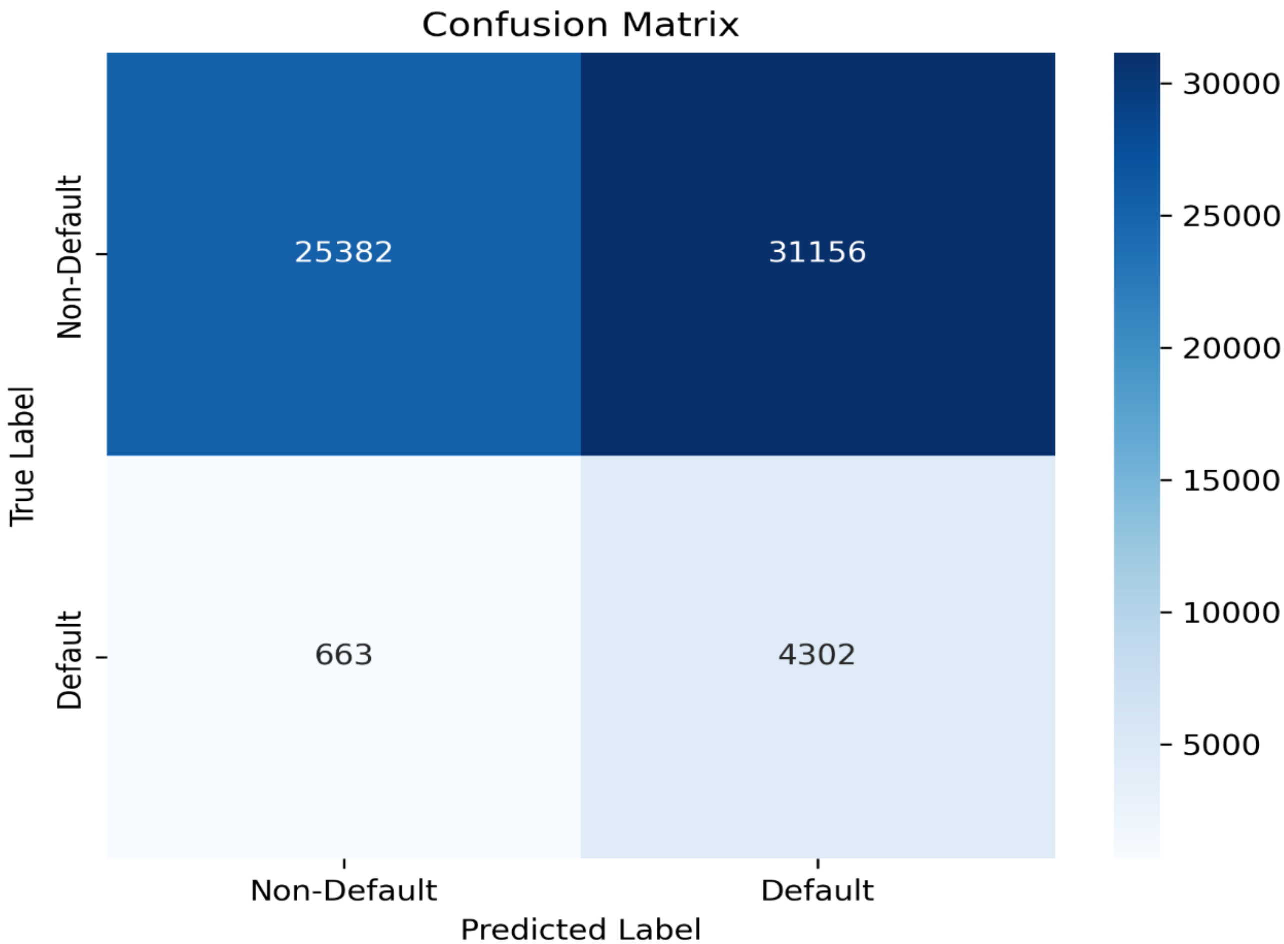

Figure 4.2: The confusion matrix of the loan default prediction model

In figure 4.2, there are a total of four types of predictions made by my model; true negatives (non-defaults predicted correctly), false positives (non-defaults incorrectly predicted to be defaults), false negatives (defaults incorrectly predicted to be non-defaults), and true positives (defaults predicted correctly). The model is able to classify 4,302 of its total 4,965 predicted defaults correctly (high recall rate) giving the model a very good ability to identify actual defaults. However, this model also predicted incorrectly by classifying many non-default borrowers to be likely to default which is typical behavior of a model functioning in the realm of imbalanced classification where finding all the potential defaults is more important than classifying more correctly (of the actual number of defaults that exist).

4.3.2 ROC Curve and AUC

The ROC (Receiver Operating Characteristic) Curve is a graphical method for assessing the performance of a binary classification model e.g. the XGBoost model that predicts if a borrower will default (positive class) or not (negative class). Figure 4.3 illustrates the Receiver Operating Characteristic curve's graphical representation of the data.

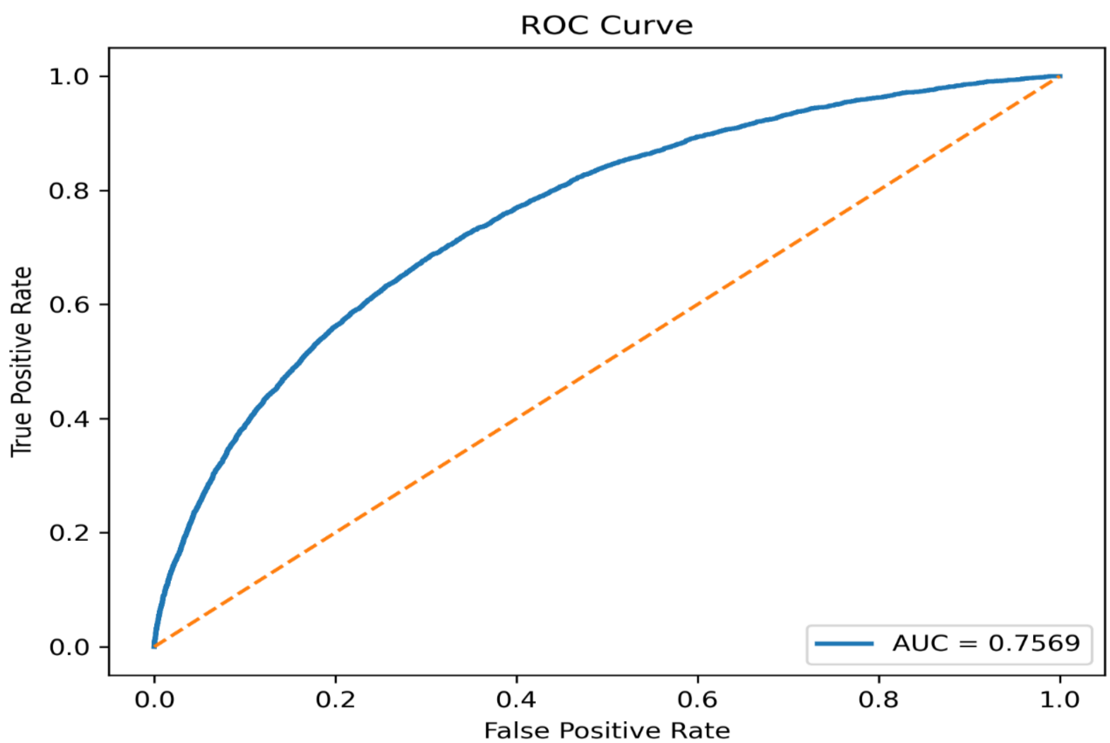

Figure 4.3 Receiver Operating Characteristic Curve

In turn, the curve is used to further assess the discriminative capability of the model. The model produced an Area Under the Curve (AUC) of 0.7569, suggesting that it has the ability to distinguish between those who will default on their loan versus those who won't at a moderate to a good level of accuracy. Generally speaking, AUCs greater than 0.75 will indicate that the model has prediction capabilities that are greater than resorting to randomly guessing, but there is still opportunity for enhancement through feature engineering, class balancing methodologies, or hyper-parameter optimization.

4.4 Result Interpretation and Implications

The analysis suggests that the XGBoost model is moderately successful in identifying high-risk borrowers for rural cooperative lending. This model excels at identifying most real defaults and therefore is instrumental in reducing financial loss; however, there are many false positives which may discourage creditworthy members from obtaining loans, ultimately limiting the availability of loans to those that are deserving.

4.5 USER INTERFACE

This section present the user interfaces for different sections of my loan prediction application, including the landing page, sign-in page, sign-up page, session page, prediction page, etc.

A. Landing page

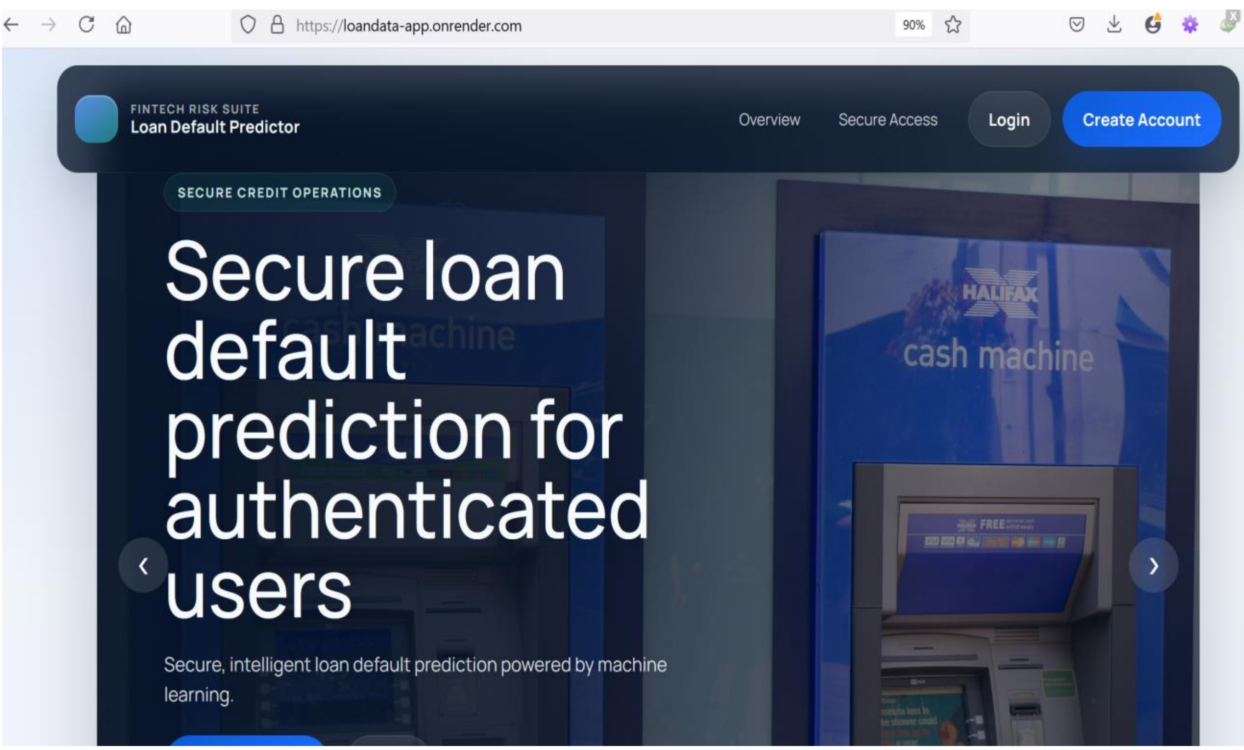

Figure 4.4 Landing page

B. Create Account Page

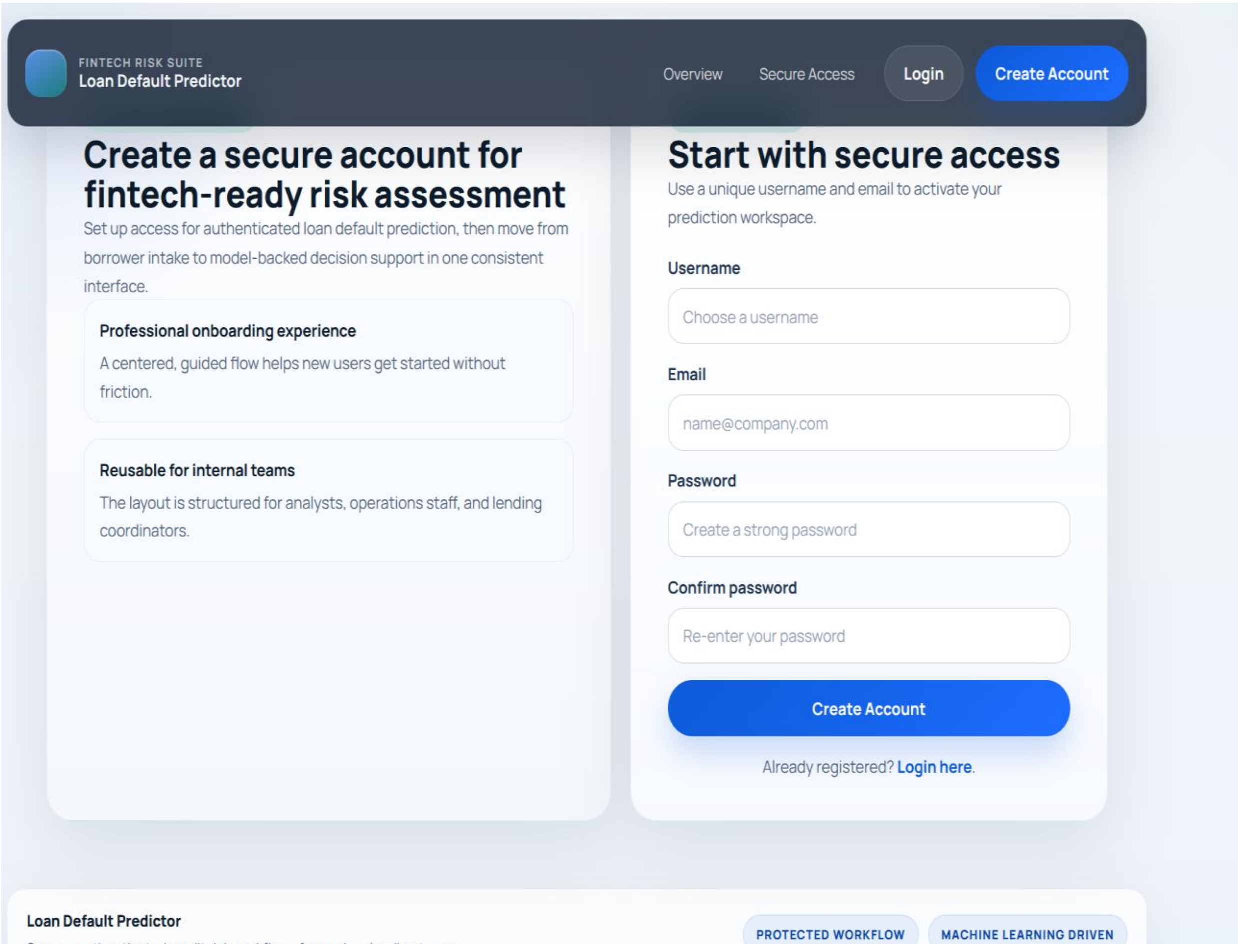

Figure 4.5 Create account page

Figure 4.5 shows the page where an intending user will need to create an account before he or she can access the loan prediction software application system.

C. Sign-in page:

A Sign-in page is a form which is designed to accept a user who have previously created an account with the software application system for loan default prediction. Figure 4.6 presents the sign in page.

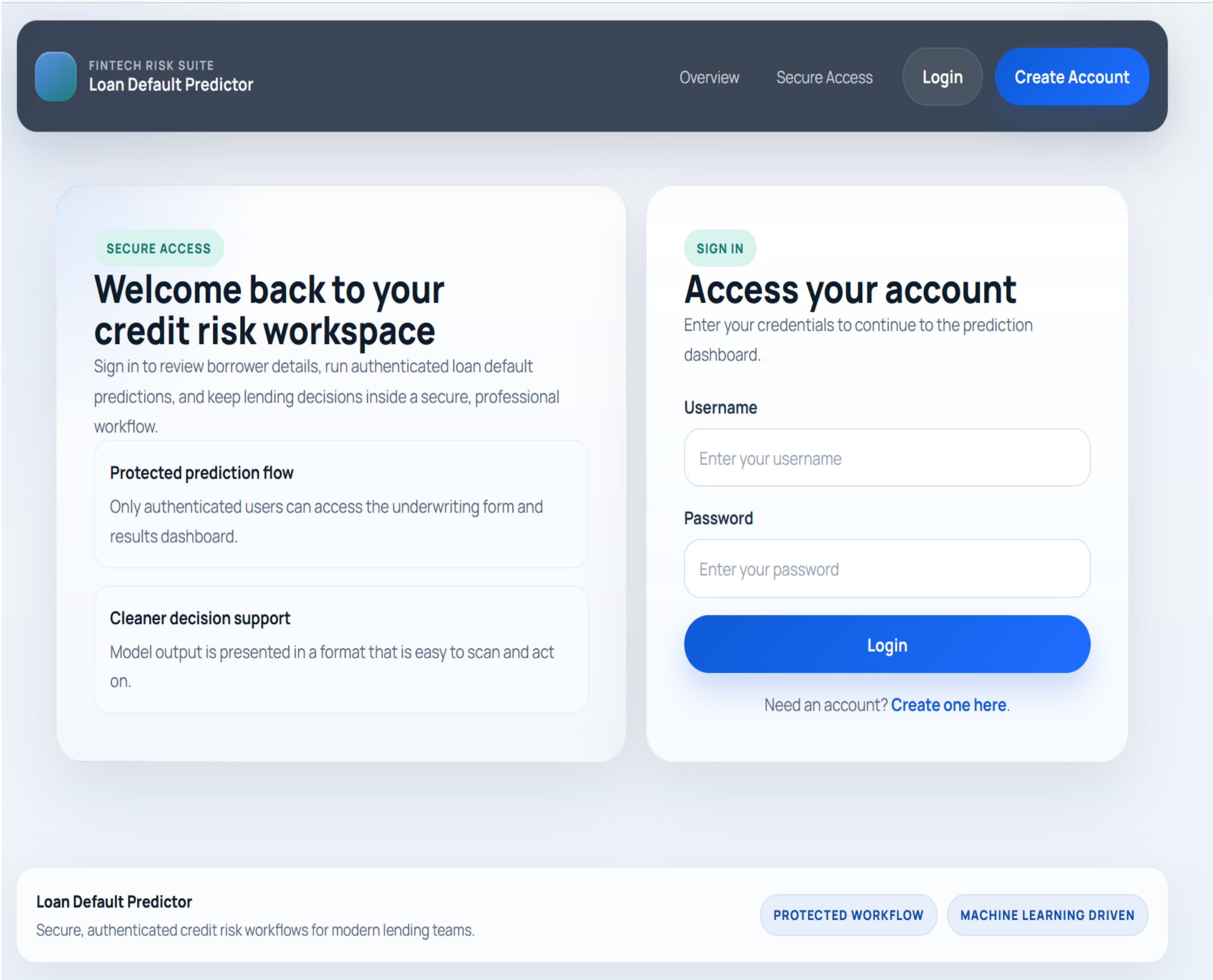

Figure 4.6 Sign-in page

D. Prediction Page

This is where the user inputs the details needed such as personal and income information, the required loan amount, and the loan's purpose.

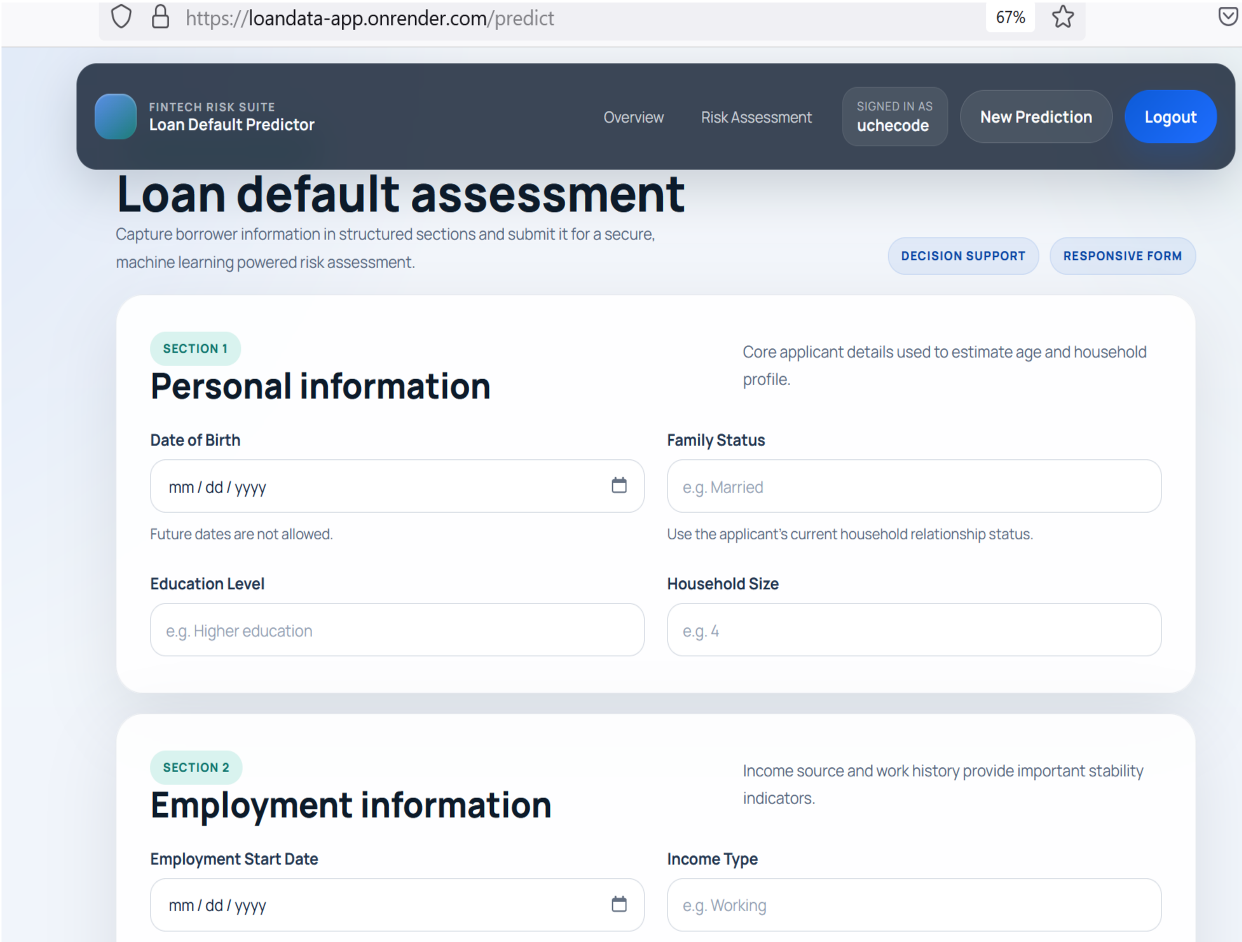

Figure 4.7 Prediction page

E. Prediction Result Page

This page displays the result of the predictions, which contains information as to whether the borrower will default and also the risk probability.

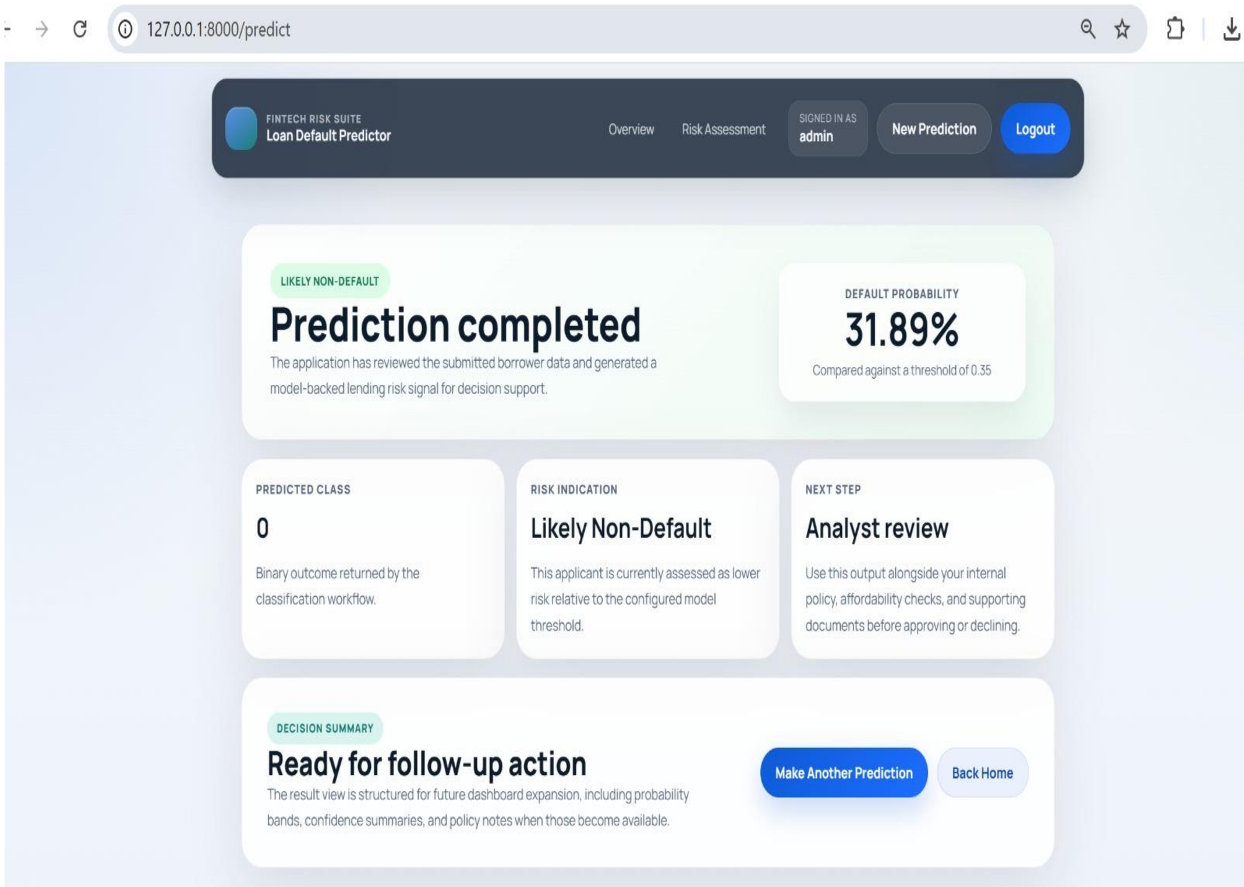

Figure 4.8 Prediction result page

CHAPTER FIVE

SUMMARY, CONCLUSION, AND RECOMMENDATIONS

5.1 Summary

This study developed a machine learning model for predicting loan default among members of rural cooperative societies using the XGBoost algorithm. Chapter One introduced the background of the study, highlighted the challenges of credit risk management in rural cooperatives, stated the problem, objectives, scope, significance, and limitations of the research.

Chapter Two is a literature review of related work in the credit risk prediction and role of machine learning in financial lending. From the review, it was observed that Gradient Boosting algorithm models, specifically XGBoost, perform consistently well when predicting loan default. It also found that the models are not being applied to cooperative societies in rural areas in Nigeria due to the different data patterns and borrower characteristics in a co-operative society as compared to a conventional bank. A research methodology was presented in Chapter Three.

Structured system analysis and design approach was used in the study. The data was preprocessed, feature engineered and selected extensively to match the context of a rural cooperative, and relevant borrower and loan features were extracted from the Home Credit data. Model implementation and evaluation were the theme of Chapter Four. The preprocessed dataset was then narrowed down to 33 features that are relevant to the task. Accuracy, confusion matrix and ROC curve were used to train and evaluate the XGBoost classifier. The model's accuracy calculated as 85.4%, and the AUC was 0.7569. Due to the class imbalance, the confusion matrix and ROC curve gave more insights on the model's capability to detect the potential defaulters.

5.2 Conclusion

The main objective of this study was to create a successful machine learning model for the prediction of loan default in a rural cooperative society using the XGBoost algorithm. The study successfully established a localized credit risk prediction system where the unique nature of the rural borrowers was taken into account by introducing variables like income, loan amount, repayment obligations, occupation type etc., and household size. The model was trained on a processed dataset, which incorporates meaningful data preprocessing and feature engineering to represent actual rural lending situations. The accuracy was 85.4% and the AUC of the model was 0.7569, indicating moderate to good discriminative power. More significantly, it was found to possess good active identification capacity of real defaulters, which is particularly important in minimizing the financial losses in cooperative societies. In conclusion, this research has made a significant impact on the use of machine learning in the financial industry, particularly in enhancing credit risk analysis for rural communities in Nigeria that are not adequately served by traditional financial institutions. It serves as a basis for making more intelligent and data-driven decisions on lending in cooperative societies.

5.3 Recommendations

The developed XGBoost model is a decent baseline model for loan default prediction, but there is some potential to further improve it. Here are some future recommendations:

1. **Hyperparameter Tuning and Optimization**: Further hyperparameter tuning and optimization of XGBoost parameters and advanced cross-validation techniques can enhance its accuracy and generalization.
2. **Handling Class Imbalance**: Investigate ways of dealing with class imbalance, including SMOTE, class weighting, and undersampling, for better handling of the imbalance between

default and nondefault cases, which may lead to higher precision and lower false positive rate.

3. **Incorporation of Additional Features**: Future studies may be able to incorporate additional, more relevant features such as repayment history patterns, group guarantee information, seasonal income fluctuations, and local economic indicators that are specific to rural communities.
4. **Deployment into a Mobile or Web Application**: Creating a mobile or web application that is easy to use would enable the cooperative officers to enter borrower information and get immediate default risk predictions that would be more useful to them in their day to day work.

References

Monje, L., Carrasco, R. A., & Sánchez-Montañés, M. (2025). Machine learning XAI for early loan default prediction. *Computational Economics*. Advance online publication. https://doi.org/10.1007/s10614-025-10962-9

Akinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O., & Ogunleye, B. (2024). Ensemble-based machine learning algorithm for loan default risk prediction. *Mathematics*, **12**, 3423. https://doi.org/10.3390/math12213423

Luo, C. (2025). *BDT-based credit default prediction*. In Proceedings of the 9th International Conference on Economic Management and Green Development. https://doi.org/10.54254/2754-1169/2025.LH23995

Özkurt, C. (2024). Enhancing financial decision-making: Predictive modeling for personal loan eligibility with gradient boosting, XGBoost, and AdaBoost. *Artificial Intelligence and Data Science Application and Research Center Journal*, **1**(1), 7–13. https://doi.org/10.69882/adba.iteb.2024072

Zhou, G., & Wang, S. (2025). Enhancing credit risk decision-making in supply chain finance with an interpretable machine learning model. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3530433

Muddu, G., Ganiyu, S. O., Ejidokun, A. O., & Aleshinloye, Y. A. (2026). *Integrated data-driven credit default prediction in Uganda using machine learning models*. Journal of the Nigerian Society of Physical Sciences. https://doi.org/10.46481/jnsps.2026.2649

Alvi, J., Arif, I., & Nizam, K. (2024). Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management. *Heliyon, 10*(24), e39770. https://doi.org/10.1016/j.heliyon.2024.e39770

Odegua, R. (2019). *Predicting bank loan default with extreme gradient boosting*. Department of Computer Science, Ambrose Alli University.

Hota, L., Jain, P. K., & Kumar, A. (2025). A comparative performance assessment for the prediction of loan approval in the financial sector. In *Advancing Financial Resilience: International Conference on Machine Learning and Data Engineering*. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2025.04.267

Puspitasari, D. M., Chin, J., & Dasman, S. (2025). Determinants and prediction model for rural bank sustainability in Indonesia post-COVID-19. *Sustainability, 17*(16), 7207. https://doi.org/10.3390/su17167207

Babarinde, T. O., Ekechi, C. C., Popoola, T. O., Oguntoye, G. A., Ayittey, S., & Ozo-ogueji, P. C. (2025). Machine learning for financial inclusion in agriculture: A study of AI-based credit scoring tools in rural Nigeria. *World Journal of Advanced Research and Reviews, 27*(2), 461–470. https://doi.org/10.30574/wjarr.2025.27.2.2884

Lai, L.-H. (2020). Loan default prediction with machine learning techniques. *2020 International Conference on Computer Communication and Network Security (CCNS)*, 5–9. https://doi.org/10.1109/CCNS50700.2020.00011

Ochieng, J. (2025). Performance of different machine learning algorithms for credit risk assessment in microfinance institutions. *Scientific Research Publishing*. https://www.scirp.org/journal/paperinformation?paperid=146944

Bhatore, S., Mohan, L., & Reddy, Y. R. (2022). Machine learning techniques for credit risk evaluation. *Journal of Discrete Mathematical Sciences and Cryptography*. https://www.semanticscholar.org/paper/Machine-learning-techniques-for-credit-risk-a-Bhatore-Mohan/2cce3461b677fa5573826ddc1cc11e12f4eb0b4d

Malhotra, R., & Malhotra, D. K. (2024). Loan default prediction using machine learning algorithms: A systematic literature review. *Pakistan Journal of Life and Social Sciences*. https://pjlss.edu.pk/pdf_files/2024_2/6234-6253.pdf

Ganawah, M. I. J., & Hadji, S. B. (2025). Impact of microfinance on rural women's income and vulnerability: A case study of Tombo community in Sierra Leone. *International Journal of Social Science and Management Research*. https://doi.org/10.37602/IJSSMR.2025.8506

Modani, A. (2025). Microfinance in different countries: Comparative analysis of different countries. *International Journal of Advanced Research, 13*(3), 1362–1374. https://doi.org/10.21474/IJAR01/20690

Sutariya, J., & Arora, G. (2023). Empowering rural economies: The role of microfinance institutions in transforming the financial landscape of rural communities. *Educational Administration: Theory and Practice, 29*(3), 1517–1522. https://doi.org/10.53555/kuey.v29i3.9991

Advans International. (2026). *The challenges of rural financial inclusion in Africa today*. https://advansinternational.com/en/insights/the-challenges-of-rural-financial-inclusion-in-africa-today/

Ramadevi, N., Kumar, V. S., Reddy, K. D., Viswanath, G. K., Sai, N. K., & Kumar, B. V. D. (2025). *EDA: Bank loan default risk analysis*. https://doi.org/10.5220/0013909100004919

Patel, B. N., & Tirthani, T. C. (2025). Impact of cooperatives on financial inclusion and livelihood in rural areas of Rajkot District, Gujarat. *International Journal of Humanities and Social Science Invention, 14*(9), 6–18. https://doi.org/10.35629/7722-14090618

Odegua, R. (2019). *Predicting bank loan default with extreme gradient boosting*. Department of Computer Science, Ambrose Alli University.

Azeez, B. A., & Ajayi, I. E. (2023). The effect of loan default on the growth of cooperative societies in Nigeria. *Nigerian Journal of Banking and Financial Issues*, **9**(1), 45–58. https://bfjournal.eksu.edu.ng/wp-content/uploads/2023/07/loan-default.pdf

Adamu, S. (2020). Effect of bank characteristics on loan default in selected microfinance banks in Nigeria. *International Journal of Research and Innovation in Social Science, 4*(8), 123–131. https://dx.doi.org/10.47772/IJRISS.2025.907000360

Afolabi, T. S., Obamuyi, T., & Egbetunde, T. (2020). Credit risk and financial performance: Evidence from Nigerian microfinance banks. *Journal of Economics and Finance, 2*(1), 8–15.

Roy, K., Chanda, S., Mandal, C., & Dasgupta, M. (2025). An empirical analysis of challenges faced by rural bank customers in health sector loans: Assessing the impact of emerging technologies and AI risks in Kolkata (West Bengal) and Mohanpur (Jharkhand). *Journal of Neonatal Surgery, 14*(12s), 471–486.

Kiros, F. (2020). Loan default rate and its determinants among smallholder farmers. *International Journal of Economics and Finance, 12*(3), 45–57. https://doi.org/10.32479/ijefi.17820

Lu, Z., Li, H., & Wu, J. (2024). Exploring the impact of financial literacy on predicting credit default among farmers: An analysis using a hybrid machine learning model. *Journal of Behavioral and Experimental Finance, 41*, 100871. https://doi.org/10.1016/j.bir.2024.01.006

Zhiqiang, L., Hongyu, L., & Junjie, W. (2024). Severe weather and collateral practices: Evidence from rural credit markets (Working Paper No. 460). *Riksbank Working Paper Series*, 1–45. https://www.riksbank.se/globalassets/media/rapporter/working-papers/2026/no.-460-severe-weather-and-collateral-practices.pdf

Balana, B. B., & Oyeyemi, M. A. (2022). Agricultural credit constraints in smallholder farming in developing countries: Evidence from Nigeria. *World Development Sustainability, 1*, 100012. https://doi.org/10.1016/j.wds.2022.100012

Umeaduma, C. M.G. (2023). Financial inclusion strategies for poverty reduction and economic empowerment in underbanked rural populations globally. *World Journal of Advanced Research and Reviews, 18*(1), 1263–1280. https://doi.org/10.30574/wjarr.2023.18.1.0709

Imran, M., Ul Haq, S., & Ozcatalbas, O. (2022). Role of microcredit in sustainable rural development. *IntechOpen*. https://doi.org/10.5772/intechopen.102588

Chen, B., & Xiao, J. (2025). Digital payments enhance both formal and informal access to credit for rural households: Evidence from China. *Frontiers in Sustainable Food Systems, 9*, 1676462. https://doi.org/10.3389/fsufs.2025.1676462

Lal, T. (2019). Measuring the impact of financial inclusion on rural development through cooperatives. *International Journal of Social Economics, 46*(3), 352–376. https://doi.org/10.1108/IJSE-02-2018-0057

Costa e Silva, E., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics, 47*(13–15), 2879–2894. https://doi.org/10.1080/02664763.2020.1759030

Hlongwane, R., Ramabao, K., & Mongwe, W. (2024). A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. *PLoS ONE, 19*(8), e0308718. https://doi.org/10.1371/journal.pone.0308718

Soomro, A. A., Zakariyah, H., Aftab, S. M. A., Muflehi, M., Shah, A., & Meraj, S. (2024). Loan default prediction using machine learning algorithms: A systematic literature review (2020–2023).

Pakistan Journal of Life and Social Sciences, 22(2), 6234–6253. https://doi.org/10.57239/PJLSS-2024-22.2.00469

Wu, Y., & Pan, Y. (2021). Application analysis of credit scoring of financial institutions based on machine learning model. *Security and Communication Networks*. https://doi.org/10.1155/2021/9222617

Tanya. (2025, August 11). Credit risk analysis: From traditional methods to digital & AI-driven approaches. *TrustDecision*. https://trustdecision.com/articles/credit-risk-analysis-from-traditional-methods-to-digital-ai-driven-approaches

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning. *Decision Analytics Journal, 3*, 100071. https://doi.org/10.1016/j.dajour.2022.100071

Cheng, Y. (2021). Research on credit strategy based on XGBoost algorithm and optimization problem. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1865/4/042137

Gao, J., Sun, W., Sui, X., & Farouk, A. (2021). Research on default prediction for credit card users based on XGBoost-LSTM model. *Discrete Dynamics in Nature and Society*, 2021, 1–13. https://doi.org/10.1155/2021/5080472

Hao, X., Zhang, Z., Xu, Q., Huang, G., & Wang, K. (2022). Prediction of f-CaO content in cement clinker: A novel prediction method based on LightGBM and Bayesian optimization. *Chemometrics and Intelligent Laboratory Systems, 220*. https://doi.org/10.1016/j.chemolab.2021.104461

Kang, Z. Z., Teh, S. Y., Tan, S. Y. G., & Ng, W. C. (2025). Loan default prediction using machine learning algorithms. *Journal of Informatics and Web Engineering*. https://doi.org/10.33093/jiwe.2025.4.3.14

Kumar, L. S., Pandimurugan, V., Usha, D., Guptha, M. N., & Hema, M. S. (2022). Random forest tree classification algorithm for predicating loan. *Materials Today: Proceedings, 57*, 2216–2222. https://doi.org/10.1016/j.matpr.2021.12.322

Liu, J., Gao, Y., & Hu, F. (2021). A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM. *Computers & Security, 106*. https://doi.org/10.1016/j.cose.2021.102289

Nkambule, D. S., Twala, B., & Pretorius, J. H. C. (2024). Effective machine learning techniques for dealing with poor credit data. *Risks, 12*(11). https://doi.org/10.3390/risks12110172

Setiawana, N., Suharjitoa, & Dianab. (2019). A comparison of prediction methods for credit default on peer-to-peer lending using machine learning. In *Proceedings of the 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019)* (Vol. 157, pp. 38–45).

Tay, W.-W., Chong, S.-C., & Chong, L.-Y. (2024). DDoS attack detection with machine learning. *Journal of Informatics and Web Engineering, 3*(3), 190–207. https://doi.org/10.33093/jiwe.2024.3.3.12

Wisdom, D. D., Vincent, O. R., Igulu, K. T., Arowolo, M. O., Hyacinth, E. A., Christian, A. U., Oduntan, O. E., & Baba, G. A. (2023). Mitigating cyber threats in healthcare systems: The role of artificial intelligence and machine learning. In *Artificial intelligence and blockchain technology in modern telehealth systems*. Institute of Engineering and Technology (IET), UK.

Hossain, M. M., Mamun, M., Munir, A., Rahman, M. M., & Chowdhury, S. H. (2025). A secure bank loan prediction system by bridging differential privacy and explainable machine learning. *Electronics, 14*(8), 1691. https://doi.org/10.3390/electronics14081691

Zhang, X., Zhang, T., Hou, L., Liu, X., Guo, Z., Tian, Y., & Liu, Y. (2025). Data-driven loan default prediction: A machine learning approach for enhancing business process management. *Systems, 13*(7), 581. https://doi.org/10.3390/systems13070581

APPENDIX A.
Code

```
 ===========================
# STEP 1: IMPORT LIBRARIES
# ===========================
import pandas as pd
import numpy as np
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt
import seaborn as sns
import joblib

from sklearn.model_selection import train_test_split
from sklearn.metrics import (
    accuracy_score,
    classification_report,
    confusion_matrix,
    roc_curve,
    auc
)
from xgboost import XGBClassifier

# ===========================
# STEP 2: LOAD DATASET
# ===========================
df = pd.read_csv("application_train.csv")

print("Initial Shape:", df.shape)

# ===========================
# STEP 3: REMOVE USELESS COLUMNS
# ===========================
if 'SK_ID_CURR' in df.columns:
    df.drop(columns=['SK_ID_CURR'], inplace=True)

# ===========================
# STEP 4: REMOVE HIGH MISSING COLUMNS
# ===========================
threshold = len(df) * 0.5
df = df.dropna(thresh=threshold, axis=1)

print("After dropping high-missing columns:", df.shape)

# ===========================
# STEP 5: SELECT IMPORTANT FEATURES
```

```
# ==========================
selected_columns = [
    'AMT_INCOME_TOTAL',
    'AMT_CREDIT',
    'AMT_ANNUITY',
    'AMT_GOODS_PRICE',
    'DAYS_EMPLOYED',
    'DAYS_BIRTH',
    'REGION_POPULATION_RELATIVE',
    'EXT_SOURCE_2',
    'EXT_SOURCE_3',
    'NAME_INCOME_TYPE',
    'OCCUPATION_TYPE',
    'NAME_EDUCATION_TYPE',
    'NAME_FAMILY_STATUS',
    'CNT_FAM_MEMBERS',
    'TARGET'
]

# Keep only columns that still exist after dropping missing-heavy columns
selected_columns = [col for col in selected_columns if col in df.columns]
df = df[selected_columns].copy()

print("After selecting important features:", df.shape)

# ==========================
# STEP 6: HANDLE MISSING VALUES
# ==========================
num_cols = df.select_dtypes(include=['int64', 'float64']).columns
df[num_cols] = df[num_cols].fillna(df[num_cols].median())

cat_cols = df.select_dtypes(include=['object', 'string']).columns
for col in cat_cols:
    mode_value = df[col].mode()[0]
    df[col] = df[col].fillna(mode_value)

# ==========================
# STEP 7: RENAME COLUMNS
# ==========================
rename_map = {
    'AMT_INCOME_TOTAL': 'income',
    'AMT_CREDIT': 'loan_amount',
    'AMT_ANNUITY': 'repayment_amount',
    'AMT_GOODS_PRICE': 'goods_price',
    'DAYS_EMPLOYED': 'days_employed',
    'DAYS_BIRTH': 'days_birth',
```

```
    'REGION_POPULATION_RELATIVE': 'region_population',
    'EXT_SOURCE_2': 'ext_source_2',
    'EXT_SOURCE_3': 'ext_source_3',
    'NAME_INCOME_TYPE': 'income_type',
    'OCCUPATION_TYPE': 'occupation',
    'NAME_EDUCATION_TYPE': 'education_type',
    'NAME_FAMILY_STATUS': 'family_status',
    'CNT_FAM_MEMBERS': 'household_size',
    'TARGET': 'loan_default'
}
df.rename(columns=rename_map, inplace=True)

# ==========================
# STEP 8: FEATURE ENGINEERING
# ==========================
df['loan_to_income_ratio'] = df['loan_amount'] / df['income'].replace(0, 1)
df['annuity_to_income_ratio'] = df['repayment_amount'] / df['income'].replace(0, 1)
df['credit_term'] = df['loan_amount'] / df['repayment_amount'].replace(0, 1)
df['age_years'] = abs(df['days_birth']) / 365
df['employment_years'] = abs(df['days_employed']) / 365

# Replace unusual employment placeholder values if present
df['employment_years'] = df['employment_years'].replace(365243 / 365, 0)

# Simulated but rule-based cooperative membership proxy
df['cooperative_member'] = np.where(
    (df['income'] < 200000) & (df['household_size'] >= 3),
    1, 0
)

# Income stability category
def income_category(x):
    if x < 50000:
        return 'low'
    elif x < 150000:
        return 'medium'
    else:
        return 'high'

df['income_stability'] = df['income'].apply(income_category)

# ==========================
# STEP 9: ENCODE CATEGORICAL VARIABLES
# ==========================
df = pd.get_dummies(df, drop_first=True)
```

```
# Convert boolean columns to integers
bool_cols = df.select_dtypes(include=['bool']).columns
df[bool_cols] = df[bool_cols].astype(int)

print("Final Shape after encoding:", df.shape)
print(df.head())

# ==========================
# STEP 10: SPLIT DATA
# ==========================
X = df.drop('loan_default', axis=1)
y = df['loan_default']

X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=42,
    stratify=y
)

# ==========================
# STEP 11: HANDLE CLASS IMBALANCE
# ==========================
ratio = (y_train == 0).sum() / (y_train == 1).sum()
print("Training Imbalance Ratio:", ratio)

# ==========================
# STEP 12: TRAIN XGBOOST MODEL
# ==========================
model = XGBClassifier(
    n_estimators=300,
    max_depth=4,
    learning_rate=0.05,
    subsample=0.8,
    colsample_bytree=0.8,
    min_child_weight=3,
    gamma=1,
    scale_pos_weight=ratio,
    random_state=42,
    eval_metric='logloss'
)

model.fit(X_train, y_train)

# ==========================
# STEP 13: PREDICTION
```

```
# ==========================
y_probs = model.predict_proba(X_test)[:, 1]

# Adjust threshold for imbalanced classification
threshold_value = 0.35
y_pred = (y_probs >= threshold_value).astype(int)

# ==========================
# STEP 14: EVALUATION
# ==========================
accuracy = accuracy_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)
fpr, tpr, _ = roc_curve(y_test, y_probs)
roc_auc = auc(fpr, tpr)

print("\nAccuracy:", accuracy)
print("\nClassification Report:\n")
print(classification_report(y_test, y_pred, zero_division=0))
print("Confusion Matrix:\n", cm)
print("AUC Score:", roc_auc)

# ==========================
# STEP 15: SAVE MODEL BUNDLE
# ==========================
bundle = {
    "model": model,
    "columns": X_train.columns.tolist(),
    "threshold": threshold_value
}
joblib.dump(bundle, "loan_default_bundle.pkl")
print("Model bundle saved as loan_default_bundle.pkl")

# ==========================
# STEP 16: CONFUSION MATRIX PLOT
# ==========================
plt.figure(figsize=(6, 5))
sns.heatmap(
    cm,
    annot=True,
    fmt='d',
    cmap='Blues',
    xticklabels=['Non-Default', 'Default'],
    yticklabels=['Non-Default', 'Default']
)
plt.title("Confusion Matrix")
plt.xlabel("Predicted Label")
```

```
plt.ylabel("True Label")
plt.tight_layout()
plt.savefig("confusion_matrix.png", dpi=300, bbox_inches="tight")
plt.close()

# =========================
# STEP 17: ROC CURVE
# =========================
plt.figure(figsize=(6, 5))
plt.plot(fpr, tpr, linewidth=2, label=f"AUC = {roc_auc:.4f}")
plt.plot([0, 1], [0, 1], linestyle='--')
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve")
plt.legend(loc="lower right")
plt.tight_layout()
plt.savefig("roc_curve.png", dpi=300, bbox_inches="tight")
plt.close()

# =========================
# STEP 18: ACCURACY CHART
# =========================
plt.figure(figsize=(5, 5))
plt.bar(["Accuracy"], [accuracy])
plt.ylim(0, 1)
plt.ylabel("Score")
plt.title("Model Accuracy")
plt.tight_layout()
plt.savefig("accuracy_chart.png", dpi=300, bbox_inches="tight")
plt.close()

print("Plots saved successfully.")
print("Generated files:")
print("- loan_default_bundle.pkl")
print("- confusion_matrix.png")
print("- roc_curve.png")
print("- accuracy_chart.png")
```