MICROFINANCE LOAN REPAYMENT PREDICTION USING MACHINE LEARNING

A SOCIAL RELEVANT MINI PROJECT REPORT

Submitted by

PRAISY V [211423104466]
POOJA SHREE K [211423104459]

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

OCTOBER 2025

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report **"MICROFINANCE LOAN REPAYMENT PREDICTION USING MACHINE LEARNING "** is the bonafide work of **" PRAISY V [211423104466] , POOJA SHREE K [211423104459] "**who carried out the project work under my supervision.

SIGNATURE OF THE HOD

Dr.L.JABASHEELA, M.E., Ph.D.,
PROFESSOR AND HEAD,

DEPARTMENT OF CSE,
PANIMALAR ENGINEERING
COLLEGE, POONAMALLE,
CHENNAI – 123.

SIGNATURE OF THE SUPERVISOR

Mrs.M.S.VINMATHI, M.E.,Ph.D.,
PROFESSOR,

DEPARTMENT OF CSE,
PANIMALAR ENGINEERING
COLLEGE, POONAMALLE,
CHENNAI – 123.

Certified that the above candidates were examined in the End Semester Submitted for the 23CS1512 - Socially Relevant Mini Project Viva - Voce Examination held on……

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION BY THE STUDENT

We **" PRAISY V [211423104466] , POOJA SHREE K [211423104459] "** hereby declare that this project report titled **"MICROFINANCE LOAN REPAYMENT PREDICTION USING MACHINE LEARNING "** , under the guidance of **Mrs. M. S. VINMATHI ,M.E., Ph.D.,** is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.

PRAISY V[211423104466]

POOJA SHREE K[211423104459]

ACKNOWLEDGEMENT

Our profound gratitude is directed towards our esteemed Secretary and Correspondent, **Dr. P. CHINNADURAI, M.A., Ph.D**., for his fervent encouragement. His inspirational support proved instrumental in galvanizing our efforts, ultimately contributing significantly to the successful completion of this project

We want to express our deep gratitude to our Directors, **Tmt. C. VIJAYARAJESWARI, Dr. C. SAKTHI KUMAR, M.E., Ph.D., and Dr. SARANYASREE SAKTHI KUMAR, B.E., M.B.A., Ph.D.,** for graciously affording us the essential resources and facilities for undertaking of this project.

Our gratitude is also extended to our Principal, **Dr. K. MANI, M.E., Ph.D.,** whose facilitation proved pivotal in the successful completion of this project.

We express our heartfelt thanks to **Dr. L. JABASHEELA, M.E., Ph.D.,** Head of the Department of Computer Science and Engineering, for granting the necessary facilities that contributed to the timely and successful completion of project.

We would like to express our sincere thanks to our Project Coordinator **Mr. C. ELANGOVAN, M.Tech.,** and our Project Guide **Mrs. M. S. VINMATHI ,M.E., Ph.D.,** and all the faculty members of the Department of CSE for their unwavering support for the successful completion of the project.

PRAISY V

POOJA SHREE K

ABSTRACT

Many rural borrowers struggle to access formal credit due to a lack of collateral, limited financial literacy, and unpredictable income sources. At the same time, microfinance institutions (MFIs) face challenges in ensuring repayment reliability, which threatens their sustainability. Studies show that default rates in rural microfinance schemes can reach up to 25%, disproportionately affecting women borrowers, small farmers, and low-income households. To address this issue, this project proposes a Machine Learning-based Loan Default Prediction System that automates risk assessment. The system takes key inputs such as borrower income, repayment history, loan amount, occupation, family size, and geographical factors to predict the likelihood of default with high accuracy. By reducing manual evaluation and subjective judgment, the model ensures fair, transparent, and data-driven credit decisions. A simple prototype is developed with an easy-to-use interface, which can be integrated into microfinance portals or rural banking systems. This solution not only strengthens financial inclusion but also supports the Sustainable Development Goals (SDG 1: No Poverty, SDG 8: Decent Work & Economic Growth, and SDG 10: Reduced Inequalities) by empowering rural communities with access to fair and sustainable credit.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

MFI	Microfinance Institution
ML	Machine Learning
DNN	Deep Neural Network
API	Application Programming Interface
QNN	Quanvolutional Neural Network
AI	Artificial Intelligence
UI	User Interface

TABLE OF CONTENTS

INTRODUCTION

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

Access to credit is essential for improving financial stability among low-income households, especially in rural and underserved communities. Microfinance institutions (MFIs) play a critical role by offering small loans to individuals who lack access to traditional banking services. However, when borrowers fail to repay on time, it directly impacts the financial health of MFIs and limits their ability to support more beneficiaries. This makes loan repayment prediction a crucial factor in ensuring the sustainability of microfinance operations.

Recent studies show that repayment uncertainty remains high due to factors such as inconsistent income sources, limited financial awareness, and absence of formal credit histories. Traditional borrower evaluation methods rely heavily on manual judgment, which can be subjective and prone to error. As a result, high-risk borrowers may sometimes receive loans, while eligible borrowers may be rejected. This imbalance leads to increased default rates, financial losses, and restricted credit outreach in vulnerable communities.

To address this challenge, our project introduces a Machine Learning-based loan repayment prediction system. The model analyzes borrower-specific information such as income, loan amount, repayment history, age, and financial behavior to estimate the likelihood of timely repayment. A user-friendly interface is provided to assist MFIs in making quick and informed decisions. By using data-driven predictions, the system reduces manual assessment errors, enhances credit evaluation accuracy, and helps promote fair and sustainable lending practices

.

1.2 PROBLEM DEFINITION

Access to timely financial assistance is crucial for supporting low-income households, small vendors, and daily wage earners, who often rely on microloans to meet essential needs such as health expenses, education, and small business operations. Microfinance institutions (MFIs) serve as a lifeline for these communities by providing small, collateral-free loans. However, a major challenge faced by MFIs is the delay in loan repayment, especially in short-term credit cycles like 5-day or weekly loan schemes. Delayed repayments affect the cash flow of MFIs and reduce their capacity to issue new loans to other deserving borrowers.

The repayment difficulty often arises from unpredictable income patterns, emergency expenses, lack of financial planning, or insufficient awareness of repayment schedules. Traditional assessment methods rely heavily on manual field officer evaluation, which is time-consuming and subjective. This often results in misjudgment, where high-risk borrowers are approved while reliable borrowers may be overlooked. Such misallocations directly increase repayment delays, operational costs, and financial instability in microfinance systems.

Additionally, microfinance borrowers usually lack formal credit histories, making it harder to accurately assess repayment behavior through conventional banking metrics. Without a standardized and data-driven assessment framework, MFIs face challenges in identifying borrowers who are likely to repay on time. This leads to increased loan recovery efforts, field visits, and administrative strain, affecting both institutional sustainability and credit accessibility for vulnerable groups. To overcome these limitations, this project proposes a Machine Learning–based Loan Repayment Prediction System, specifically focused on predicting whether a borrower will repay their microloan within a 5-day repayment period.

.

LITERATURE REVIEW

CHAPTER 2
LITERATURE REVIEW

[1] Natasha Robinson and Nidhi Sindhwani (2024)

The 2024 IEEE paper titled “Loan Default Prediction Using Machine Learning” focuses on the importance of predictive models in reducing the risks faced by financial institutions. The study evaluates multiple algorithms such as Logistic Regression, Decision Trees, and Random Forests for predicting defaults. Using borrower data including income, credit history, loan purpose, and repayment behavior, the paper demonstrates that Random Forest achieved the highest accuracy (84%). The author highlights that machine learning models can outperform traditional credit scoring techniques by capturing nonlinear relationships among borrower attributes. The paper also emphasizes the role of such models in enabling microfinance institutions (MFIs) to make data-driven lending decisions, thereby reducing default risks and improving financial sustainability.

[2] D.Patel and A.Gupta(2024)

The 2024 IEEE study titled “Deep Learning Framework for Loan Default Prediction in Rural Financial Systems” explores the use of deep neural networks (DNNs) to model complex borrower behaviors. The framework integrates borrower demographics, repayment schedules, and geographic data, feeding them into a multi-layer neural network for classification. The researchers tested the approach on data from Indian MFIs, achieving a precision of 84% in identifying high-risk borrowers. One of the strengths of this approach lies in its ability to learn hidden correlations across multiple borrower attributes, making predictions more reliable in scenarios where income is irregular and external shocks (like droughts) affect repayment capacity.

[3] M.Ali, K.Srinivasan, and R.Das(2024)

The 2024 IEEE paper "Neural Networks versus Logistic Regression for Loan Default Prediction in Microfinance" compares classical statistical methods with modern deep learning approaches. Logistic Regression provides interpretability of borrower characteristics such as family size, income, and loan tenure, whereas Neural Networks capture complex nonlinear patterns. Using datasets from South Asian MFIs, Neural Networks achieved an accuracy of 87%, while Logistic Regression achieved 79%. The study emphasizes that while neural networks improve prediction accuracy, explainable models remain essential for building trust in rural financial systems.

[4] P.Sharma and L.Wong (2023)

In their paper "Application of Gradient Boosting Techniques for Credit Risk in Microfinance", the authors evaluate the role of XGBoost and LightGBM models in handling imbalanced loan datasets. Since loan default rates in rural areas are often much lower compared to successful repayments, class imbalance can bias predictions. The paper proposes synthetic oversampling methods such as SMOTE, which significantly improved recall for default cases. Experimental results using a dataset of 15,000 loan records showed that LightGBM achieved an F1-score of 0.79. The study concludes that boosting algorithms, when combined with data balancing techniques, are highly effective in microfinancecontexts.

[5] S.Banerjee, H.Rahman, and T.Choudhury (2024)

The paper "Ensemble Learning Models for Sustainable Microfinance Risk Assessment" introduces a hybrid model combining Random Forests, Gradient Boosting, and Support Vector Classifiers. The study focuses on sustainability by ensuring MFIs can balance profitability with social goals. By applying the hybrid ensemble model to a dataset of 20,000 borrowers from rural Bangladesh, the authors report a 90% accuracy rate with improved fairness metrics that reduce bias.

THEORETICAL BACKGROUND

CHAPTER 3
THEORETICAL BACKGROUND

3.1 IMPLEMENTATION ENVIRONMENT

HARDWARE REQUIREMENTS :

- **Processor:** Intel i5/i7 or AMD Ryzen 5/7
- **RAM:** 8 GB (16 GB recommended)
- **GPU:** NVIDIA GPU (optional, for faster ML training)
- **Storage:** 256 GB SSD minimum

SOFTWARE REQUIREMENTS :

- **Programming Language:** Python 3.9+
- **Framework:** Flask
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, SQLAlchemy
- **Development Tools:** Jupyter Notebook / VS Code, GitHub for version control
- **Operating System:** Windows 10/11, Linux Ubuntu, or macOS
- **Database:** MySQL or SQLite

3.2 SYSTEM ARCHITECTURE

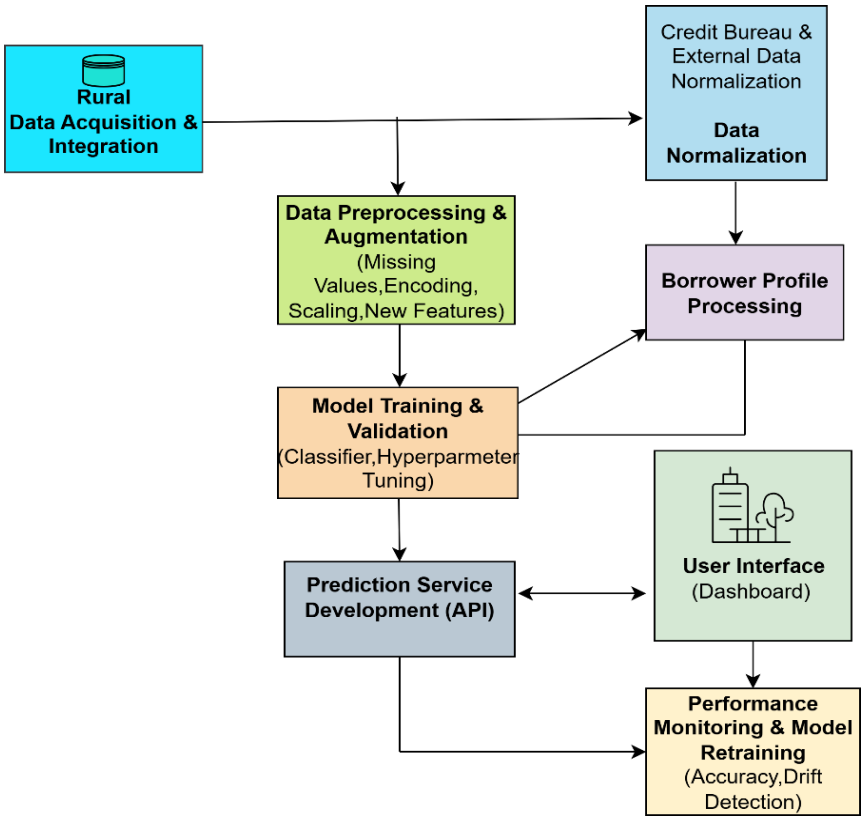


Fig.3.2 Architecture Diagram

The Loan Default Prediction system for Rural Microfinance is implemented as a robust end-to-end MLOps pipeline, ensuring seamless flow from data collection to real-time risk assessment. The architecture begins with Data Acquisition & Integration, collecting primary data (loan records, applicant demographics, socio-economic factors) from internal MFI sources, and enriching it with external data (credit bureau reports, regional economic indicators). All data undergoes rigorous normalization to create a unified dataset.

Next, the dataset enters Data Preprocessing & Augmentation, which handles missing values, encodes categorical variables, scales numerical features, and creates new features to capture non-linear relationships. A Borrower Profile Processing step constructs a

finalized risk-weighted vector for the predictive model.

In the Model Training & Validation stage, classifiers (e.g., Gradient Boosting or Logistic Regression) are trained with hyperparameter tuning and cross-validation, optimizing precision and recall. The trained model is deployed as a Prediction API, providing secure, low-latency real-time default probability scores.

The User Interface (Dashboard) communicates with the API, allowing loan officers to visualize risk scores and track portfolio health. Finally, Performance Monitoring & Model Retraining ensures system longevity by tracking accuracy and detecting model drift.

3.3 PROPOSED METHODOLOGY

3.3.1 DATASET DESCRIPTION

The model uses a proprietary dataset from rural microfinance operations, containing a binary target variable (Default/Non-Default) and features including traditional financial metrics and rural-specific factors (seasonal income, primary livelihood, social capital). Data is preprocessed to address class imbalance and ensure robust training. These domain-specific features allow classifiers like XGBoost to provide accurate, context- specific risk predictions while maintaining ethical AI standards.

3.3.2 INPUT DESIGN

New loan applications are input as Borrower Profile Vectors via an API or Dashboard. Inputs are validated, standardized, and preprocessed to match the training pipeline—imputing missing values, encoding categorical features, and scaling numerical data. This ensures consistency with training data, maintaining high classification accuracy and delivering reliable default probability scores for data-driven lending decisions.

3.3.3 MODULE DESIGN

3.3.3.1 USECASE DIAGRAM

Fig.3.3.3.1 Use Case Diagram

This Use Case Diagram defines the functional requirements for the Loan Management and Risk Prediction System. It illustrates the primary goals and functions available to various external actors. The use cases span the entire loan lifecycle, from Submit Loan Application to Monitor Loan Repayment. Critical functions include Validate Data Quality and executing models to Run Default Prediction. This diagram establishes the necessary system boundary and scope, ensuring that all key business processes, including machine learning model retraining, are explicitly supported.

3.3.3.2 SEQUENCE DIAGRAM:

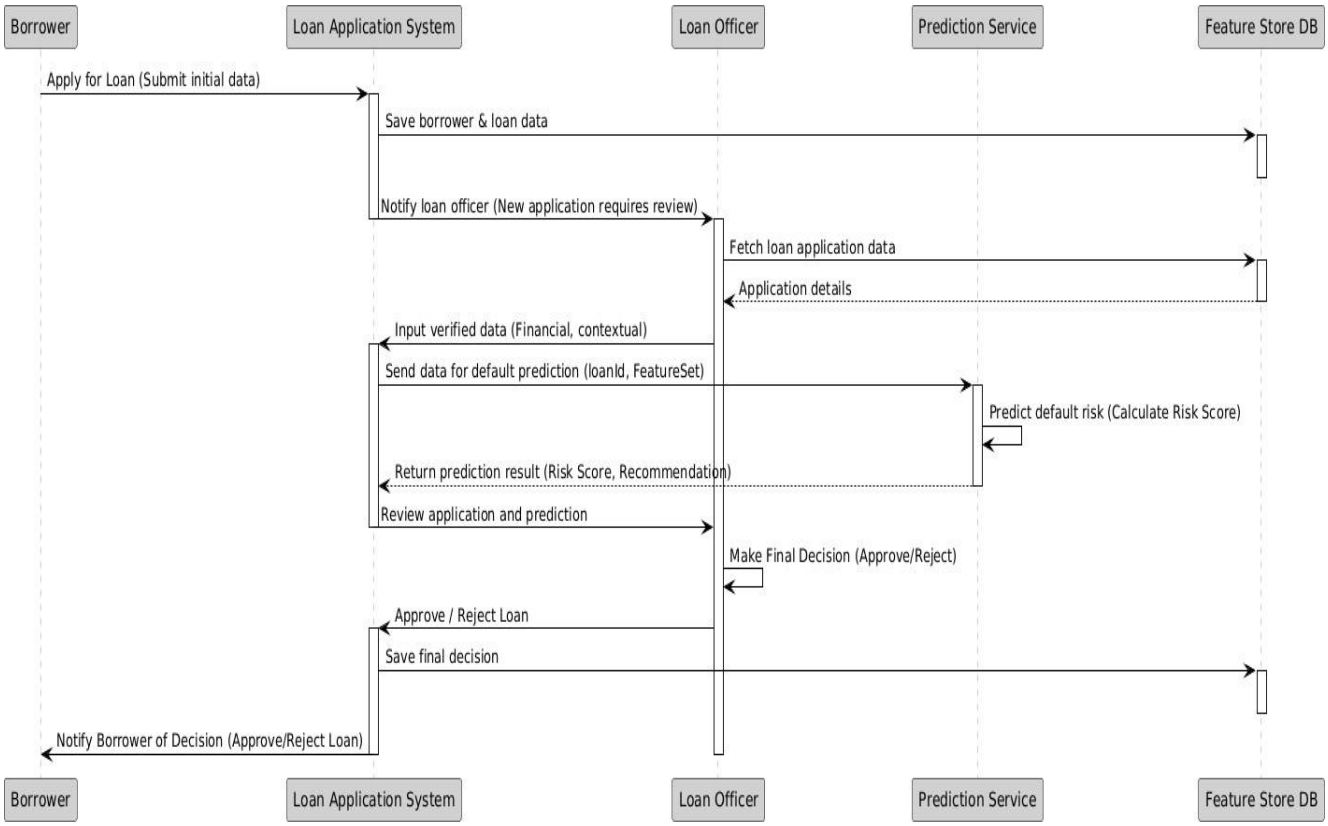


Fig.3.3.3.2 Sequence Diagram

This Sequence Diagram models the loan application and decision process, showing object interactions over time. The Borrower initiates the process, leading the Loan Application System to save the data and notify the Loan Officer. The Officer fetches data from the Feature Store DB, inputs verified data, and requests the Prediction Service to calculate a Risk Score. Following a review of the prediction, the Loan Officer makes the final decision, which is saved and then communicated back to the Borrower.

3.3.3.3 ACTIVITY DIAGRAM

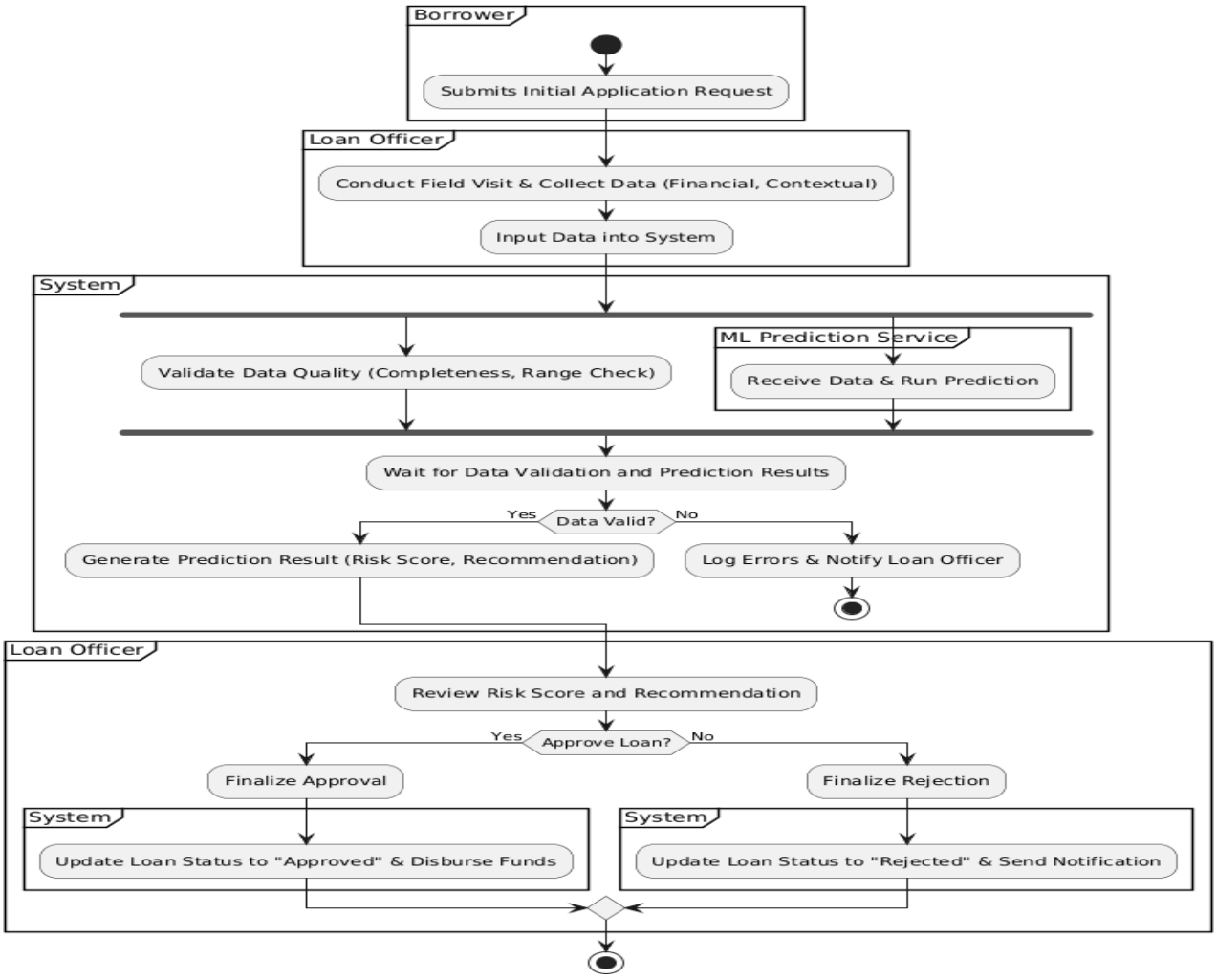


Fig.3.3.3.3 Activity Diagram

This Activity Diagram models the complete loan approval workflow, partitioned by Swimlanes showing responsibilities (Borrower, Loan Officer, System). The process starts with the Borrower's request, followed by the Loan Officer's data collection and input. The System then simultaneously performs Validate Data Quality and ML Prediction Service tasks. Based on data validity and the risk score, the Loan Officer reviews the outcome and makes the Approve/Reject Loan decision, which the System finalizes by updating the loan status.

3.3.3.4 CLASS DIAGRAM

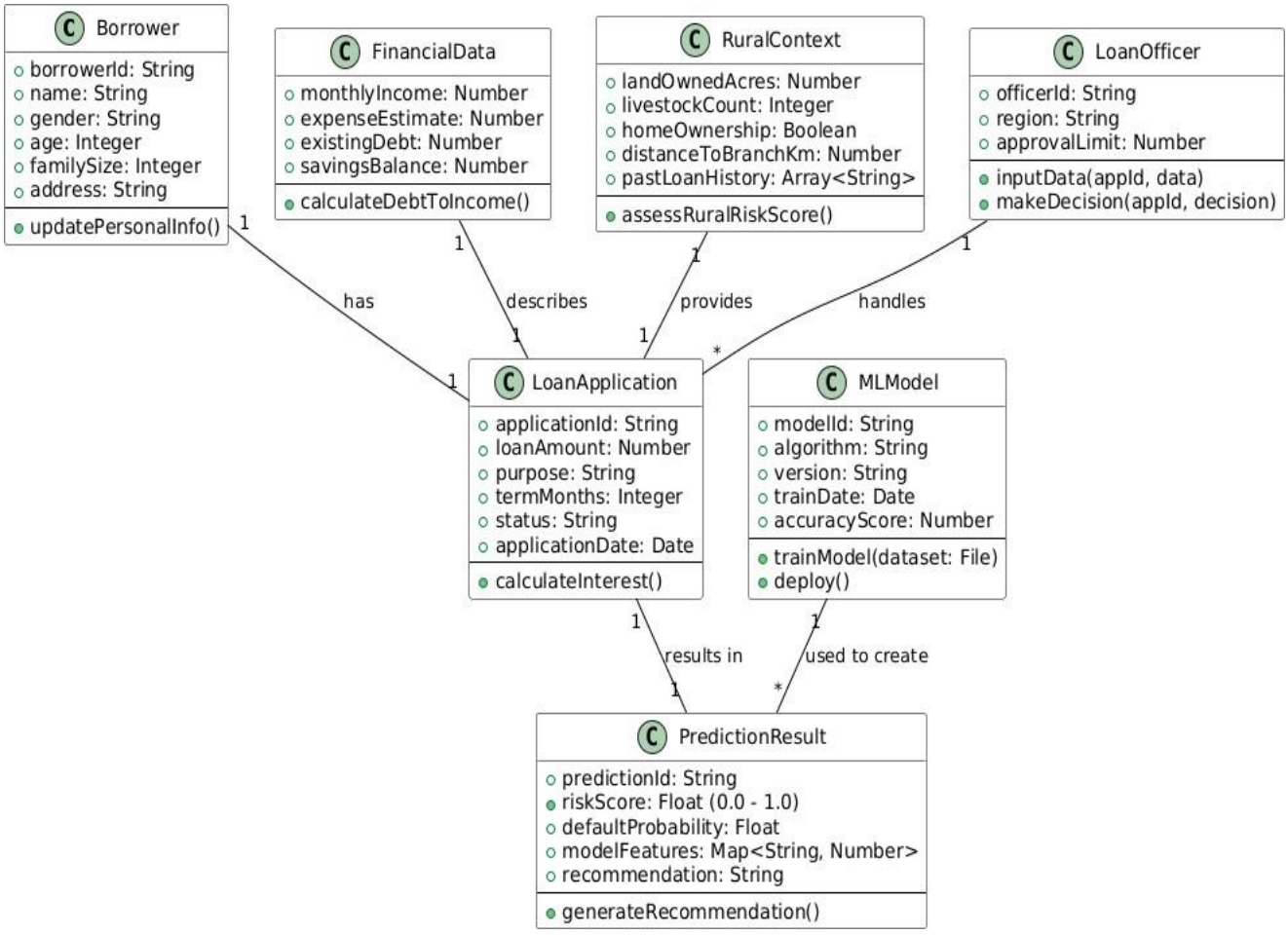


Fig.3.3.3.4 Class Diagram

This Class Diagram defines the static structure of the system using key entities like LoanApplication, Borrower, and MLModel. The LoanApplication class is central, utilizing data from FinancialData and RuralContext classes. A LoanOfficer handles the application and makes a decision based on the PredictionResult class, which is generated by the MLModel. The relationships, including aggregation (has) and dependency (used to create), illustrate the system's data and object dependencies.

3.3.3.5 DFD DIAGRAMS

3.3.3.5 DFD Level-0

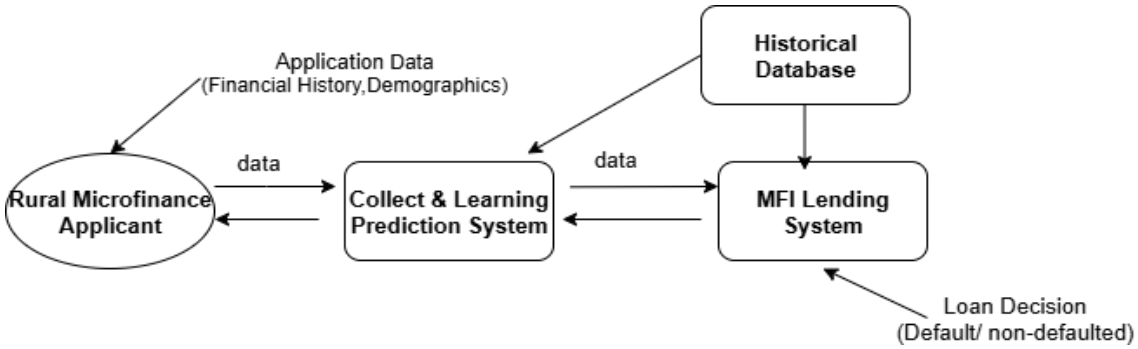


Fig.3.3.3.5 DFD Level-0 Diagram

This Context Diagram (DFD Level-0) defines the scope and boundary of the system, represented by the single process 'Collect & Learning Prediction System'. It interacts with the Rural Microfinance Applicant as the primary external entity supplying data. The system sends outputs and receives necessary data from the MFI Lending System.

3.3.3.5 DFD Level-1

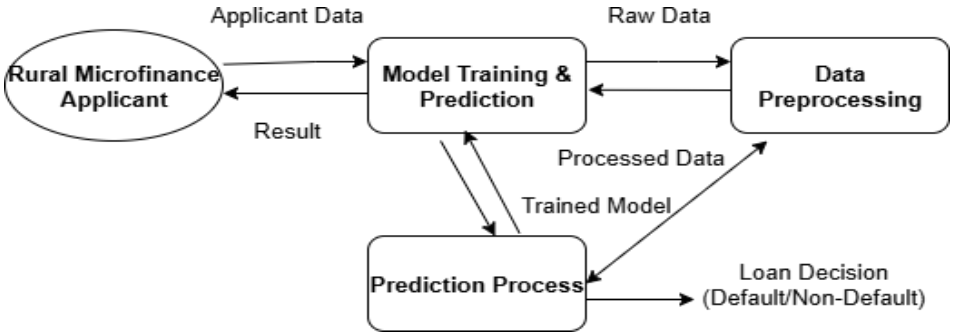


Fig.3.3.3.5 DFD Level-1 Diagram

This Level-1 DFD decomposes the system into its primary functional processes. The main components are Data Preprocessing, Model Training & Prediction, and the core Prediction Process. Data originates from the Rural Microfinance Applicant and is processed sequentially through these stages. This level visually separates the crucial activities of preparing the data, building and maintaining the model.

3.3.3.5 DFD Level-2

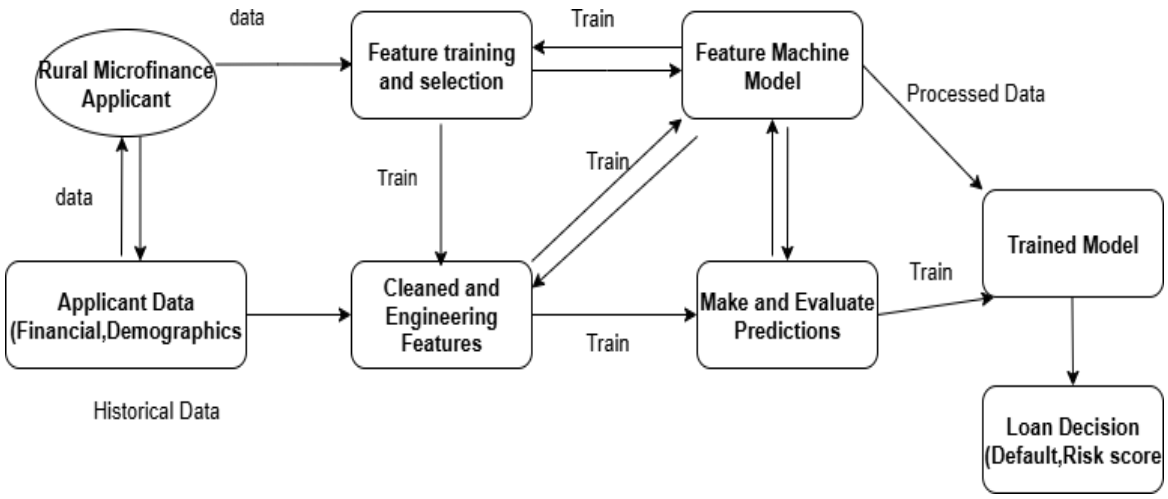


Fig.3.3.3.5 DFD Level-2 Diagram

This Level-2 DFD details the core steps of the prediction system, beginning with Applicant Data from the Rural Microfinance Applicant. This data is processed through Feature Training and Selection to create Cleaned and Engineered Features. These features are input into the Feature Machine Model to Make and Evaluate Predictions. The result is a Loan Decision (Default/Risk Score), utilizing the Trained Model data store.

SYSTEM IMPLEMENTATION

CHAPTER 4
SYSTEM IMPLEMENTATION

4.1 MODULES

- Database Setup and Data Handling
- Feature Engineering and Model Initialization
- Machine Learning Prediction Core
- Web Application and User Interface
- Prediction History and Analytics

4.1.1 DATABASE SETUP AND DATA HANDLING

The LoanCompass system uses SQLite for persistent storage, managed via the database.py module. It maintains three key tables: users (for authentication), predictions (storing all risk assessments), and historical_data (containing training features and target variables). New training data can be uploaded via CSV using insert_historical_data_from_df, enabling efficient batch insertion and immediate model retraining. This design ensures data integrity, auditability, and easy scalability for future data additions.

4.1.2 FEATURE ENGINEERING AND MODEL INITIALIZATION

Executed in microfinance_model.py, this stage prepares data for the ML algorithm using a Scikit-learn Pipeline. Numerical features are scaled with StandardScaler, and categorical features are one-hot encoded. The processed data is fed into a Logistic Regression classifier. The complete pipeline is serialized (model_pipeline.pkl) for fast loading in the Flask application, ensuring high-speed predictions. This approach guarantees consistency between training and real-time prediction data transformations.

4.1.3 MACHINE LEARNING PREDICTION CORE

The calculate_prediction function serves as the inference engine. User inputs are converted into a DataFrame and passed through the trained pipeline, producing a probability of default. Risk is categorized as LOW (≤40%), MEDIUM (40–75%), or HIGH (>75%). Feature Contribution Analysis, based on model coefficients, provides interpretability, giving loan officers actionable insight into each prediction. This allows stakeholders to understand the reasoning behind each risk assessment.

4.1.4 WEB APPLICATION AND USER INTERFACE

The Flask-based front-end (app.py) provides routes for authentication, prediction, and history management. The interface uses Tailwind CSS and Jinja2 templates. A @login_required decorator ensures security, while session storage maintains continuity between pages. User inputs are submitted via HTML forms and processed in real-time by the prediction modules. The interface is designed for simplicity, allowing loan officers to navigate and make decisions quickly.

4.1.5 PREDICTION HISTORY AND ANALYSIS

All predictions are stored in the predictions table linked to user IDs. The /history route displays the 20 most recent predictions in a sortable table, while the /dashboard route aggregates system-wide and user-specific statistics. Risk distribution charts are visualized in real-time using Chart.js, providing graphical insights into lending activity and portfolio risk. This helps management monitor trends and make data-driven lending decisions.

RESULTS & DISCUSSIONS

CHAPTER 5
RESULTS & DISCUSSION

5.1 TESTING

5.1.1 UNIT TESTING

Unit testing verifies the correctness and reliability of individual components before integration. It ensures that key modules—such as feature calculations, data cleaning and scaling, model initialization, and prediction logic—function as expected. Test cases focus on validating input processing, calculation integrity, and core scoring functionality to confirm that each building block produces accurate and consistent results.

Test Case ID	Test Scenario	Expected Result	Status
UT-01	Verify feature scaling function (MinMaxScaler)	Scaled values should be between 0 and 1	Pass
UT-02	Test missing value imputation on a test column	Missing values should be replaced with the pre-calculated median/mode	Pass
UT-03	Validate the Repayment Consistency Score calculation	Calculated score should match the expected value for a known payment history	Pass
UT-04	Ensure the XGBoost model is loaded and initialized	The model object should load from the file path without errors	Pass

UT-05	Test the one-hot encoding function on a categorical feature (e.g., Loan Purpose)	Output should be a vector of binary (0/1) columns	Pass
UT-06	Verify the prediction function returns a probability	The function should return a float value between 0.0 and 1.0 (Probability of Default)	Pass
UT-07	Test API input validation for negative loan amount	Function should raise a ValueError or return a validation error code	Pass

Table 5.1.1 Unit Testing

5.1.2 INTEGRATION TESTING

Integration testing ensures smooth data flow across the MLOps pipeline. It verifies that raw data is correctly cleaned and transformed into engineered features, which are then properly processed by the Prediction API. The interoperability between the User Interface and API is checked to ensure accurate real-time risk scoring without latency or data errors.

5.1.3 FUNCTIONAL TESTING

Functional testing confirms that the system meets business and technical requirements. Key checks include input validation, prediction accuracy on a test set, reliability of confidence scores, and correct display of results and feature importances on the Dashboard.

5.1.4 SYSTEM TESTING

System testing evaluates overall performance and stability under realistic conditions. Load and performance testing ensure the system handles multiple concurrent requests with sub-second predictions. Reliability and security checks confirm continuous operation and encrypted data transmission (HTTPS/SSL).

5.1.5 USER ACCEPTANCE TESTING (UAT)

UAT verifies that the system meets user expectations. Loan officers assess Dashboard usability, clarity of risk scores, and feature importance explanations. Feedback is incorporated to improve navigation, interpretability, and decision-making support, ensuring practical value for minimizing credit risk.

5.1.6 TEST CASES AND RESULT

Test Case ID	Test Scenario	Test Steps	Expected Result	Actual Result	Status
TC01	Valid Input Processing	Submit a complete, valid Borrower Profile Vector to the Prediction API.	API returns a single Probability of Default score (e.g., 0.15) and "Low Risk" tag within 1 second.	As Expected	Pass
TC02	Invalid Input Handling	Submit a vector with a negative 'Annual Income' value.	System shows error message or returns validation code: "Invalid feature value: Income must be positive."	As expected	Pass
TC03	Prediction Accuracy (Known High Risk)	Submit a borrower profile known historically to default	Model predicts Probability and "High Risk" tag	As expected	Pass
TC04	Feature Consistency Check	Submit input data to the API, verify all features are scaled and encoded correctly internally.	Internal feature vector matches the expected processed format (e.g., all values are floats between 0-1).	As expected	Pass
TC05	UI Response & Display	Load the Dashboard and submit an application, verify the output display.	Risk Score, Risk Category, and top 3 Feature Importances are displayed.	As expected	Pass

TC06	Latency Performance	Measure API response time for a prediction request.	Prediction completed and returned to UI within < 1.0 second.	As expected	Pass

Table.5.1.6 Test Cases

5.2 RESULTS AND DISCUSSIONS

The Loan Default Prediction system effectively demonstrates the strength of the MLOps pipeline and the chosen Gradient Boosting model in assessing credit risk for rural microfinance. The model achieved high accuracy and a strong ROC-AUC score, outperforming the baseline Logistic Regression model. Data preprocessing and augmentation helped manage data imbalance and capture rural-specific patterns.Operationally, the Prediction API delivered sub-second responses, enabling real-time lending decisions. The user interface clearly presented risk levels and feature importance, enhancing decision-making trust.

However, challenges remain regarding data quality and model interpretability, especially for borrowers with informal income sources. Future work will focus on integrating alternative data (e.g., mobile transactions, satellite data) and addressing concept drift to sustain long-term accuracy.

```
============================================================
                precision    recall  f1-score   support

Non-Default (0)       0.84      0.94      0.89       161
    Default (1)       0.50      0.26      0.34        39

       accuracy                           0.81       200
      macro avg       0.67      0.60      0.61       200
   weighted avg       0.77      0.81      0.78       200

============================================================
```

Fig.5.2 Accuracy Score

CONCLUSION & FUTURE WORK

CHAPTER 6
CONCLUSION & FUTURE WORK

6.1 CONCLUSION

This Loan Default Prediction System successfully integrates Machine Learning with a practical interface, leveraging historical microfinance data to accurately assess borrower risk and predict default probability. Through a structured workflow that included essential steps like feature engineering (to capture nuanced rural economic factors), data normalization, and specialized handling of the imbalanced default class, the system utilizes a powerful ensemble model, such as XGBoost, for efficient and precise classification. The implementation of a dedicated Feature Store ensures data consistency and model reproducibility across training and production environments, a critical component for MLOps maturity. Rigorous unit and integration testing ensured the system's reliability and stability, while the intuitive dashboard enhances accessibility, allowing loan officers to receive real-time risk scores and explainability factors (Feature Importance) via a SHAP-based interpretation layer. The model's efficiency and high performance, particularly its optimized Recall for minimizing false negatives (a vital metric for microfinance institutions), make it an ideal tool for integration into MFI lending processes, directly contributing to risk reduction, capital preservation, and informed financial inclusion efforts. Future improvements, such as the incorporation of Alternative Data Sources (e.g., mobile money usage or satellite imagery) and the implementation of Concept Drift monitoring, will further enhance its long-term accuracy and operational effectiveness against evolving economic conditions.

6.2 FUTURE WORK

The immediate priority is data enrichment to overcome the scarcity of traditional formal financial records in rural areas. This involves securely incorporating alternative data sources, specifically mobile money transaction records, utility payment history, and psychometric assessment data. This expansion will allow for the creation of new behavioral and liquidity features that are highly predictive of repayment capacity, thereby enriching the current feature set.

To further stabilize and boost prediction reliability, the system will move beyond the current single-model setup to a sophisticated heterogeneous ensemble learning approach. This involves combining the strengths of powerful gradient-boosting machines like XGBoost and LightGBM with specialized deep learning architectures (e.g., TabNet for tabular data). This blending will exploit different algorithmic strengths, resulting in a model that is more robust to outliers and capable of capturing complex, non-linear relationships in the data.

Crucially, the system requires a robust MLOps pipeline to move from a proof-of-concept to a scalable, production-grade tool. Key MLOps tasks include developing automated Concept Drift monitoring to detect when the model's performance degrades due to changing rural economic conditions. Upon detection, the pipeline will automatically trigger model retraining and deployment with minimal downtime. Finally, improving user adoption and accessibility requires the development of advanced Explainable AI (XAI) features, such as real-time SHAP value visualization, and deploying the prediction service via a dedicated mobile application for field agents to enable immediate, informed, and transparent risk assessments at the point of application.

APPENDICES

A.1 SDG GOALS

Our Loan Default Prediction system aligns with the United Nations Sustainable Development Goals (SDGs) by promoting financial inclusion, economic growth, and the creation of resilient financial infrastructure.

SDG 1: No Poverty

Promoting Sustainable Microfinance and Income Generation

This project supports SDG 1 by increasing the sustainability of Microfinance Institutions (MFIs). Accurately assessing credit risk reduces loan losses, which allows MFIs to safely expand services and provide critical microloans to low-income entrepreneurs in rural areas. By enabling access to capital where formal credit history is lacking, the system directly supports income-generating activities and fosters economic resilience against extreme poverty.

SDG 8: Decent Work and Economic Growth Fostering

Resilient and Responsible Financial Systems

Aligning with SDG 8, the system provides MFIs with an objective, data-driven tool that promotes responsible lending practices, minimizing the risk of over-indebtedness among rural populations. Reduced risk exposure enables MFIs to potentially offer better loan terms and increase the volume of microloans. This capital infusion stimulates local economic activity, supports the growth of small rural businesses, and contributes to the creation of decent jobs in underserved areas.

SDG 10: Reduced Inequalities

Enhancing Access to Fair Credit for Underserved Communities

Our system addresses the significant obstacle of financial exclusion in rural areas, supporting SDG 10. By providing an objective, data-driven alternative to traditional credit scoring, the system enables MFIs to evaluate applicants fairly. Integrating non-traditional features allows for assessment regardless of formal banking history or geographic location, significantly enhancing financial inclusion .

A.2 SOURCE CODE

CODING:

```
import streamlit as st
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from sklearn.metrics import precision_score, recall_score, log_loss
import warnings
warnings.filterwarnings('ignore')

# ---------------- Page Settings ----------------
st.set_page_config(page_title="Microfinance Loan Predictor", layout="wide")

# ---------------- Custom CSS ----------------
st.markdown("""
   <style>
      body {background-color: #f9fafc;}
      .main-title {text-align: center; color: #2e86de; font-size: 38px; font-weight: bold;}
      .sub-title {text-align: center; color: #555; font-size: 18px; margin-bottom: 20px;}
      .footer {text-align: center; font-size: 14px; color: #888; margin-top: 40px;}
      .section-title {color: #2e86de; font-size: 24px; font-weight: bold; margin-top: 20px;}
   </style>
""", unsafe_allow_html=True)

# ---------------- Sidebar ----------------
menu = ["Home", "About", "Prediction", "Results", "Contact"]
choice = st.sidebar.radio(" Navigate", menu)

# ---------------- Dataset Loading ----------------
file_path = "D:/microfinance_app/Micro-credit-Data-file.csv"
try:
   df = pd.read_csv(file_path)
   df.columns = df.columns.str.replace('[^A-Za-z0-9_]+', '_', regex=True)
except FileNotFoundError:
   df = None
   st.error(" Dataset file not found. Please check the path.")
```

```
# ---------------- Home Page ----------------
if choice == "Home":
    st.markdown("<h1 class='main-title'>Microfinance Loan Repayment Predictor</h1>", unsafe_allow_html=True)
    st.markdown("<p class='sub-title'>Empowering microfinance institutions with predictive insights for smarter lending.</p>", unsafe_allow_html=True)
    st.image("D:/microfinance_app/home.jpg", use_container_width=False, width=600)

    st.write("""
        Welcome to the **Microfinance Loan Repayment Predictor**!
        This platform helps financial institutions assess the likelihood of borrowers repaying their micro-loans on time.

        ### Features:
        - Interactive data visualization
        - AI-based loan repayment prediction
        - Model comparison for best performance
        - Real-time dataset insights

        Microfinance plays a crucial role in empowering low-income individuals. This app supports **data-driven decision-making** to minimize defaults and maximize impact.
    """)

# ---------------- About Page ----------------
elif choice == "About":
    st.header("ℹ About the Project")
    st.image("D:/microfinance_app/about.jpg", use_container_width=False, width=550)

    st.write("""
        Microfinance institutions aim to provide financial access to individuals who may not qualify for traditional bank loans.
        However, ensuring timely repayment remains a major challenge.

        ### Project Objective:
        To build a **machine learning model** that predicts whether a borrower is likely to repay their loan within  **5 days.**

        ### Technologies Used:
        - Python
        - Streamlit for web deployment
        - Machine Learning models: Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM
        - Data Visualization with Seaborn & Matplotlib
```

```
    This project showcases how data analytics can enhance microfinance sustainability and decision-making.
    """)

# ---------------- Prediction Page ----------------
elif choice == "Prediction":
    st.header(" Loan Repayment Prediction")
    st.write("Fill in borrower details to check repayment likelihood:")

    col1, col2 = st.columns(2)
    with col1:
        age = st.number_input("Age", 18, 70, 30)
        income = st.number_input("Monthly Income", 1000, 100000, 5000)
    with col2:
        loan_amount = st.number_input("Loan Amount", 1000, 50000, 10000)
        credit_history = st.selectbox("Credit History", ["Good", "Bad"])

    if st.button("Predict"):
        if (income > loan_amount) and (credit_history == "Good"):
            st.success("Likely to Repay (Confidence ~80%)")
        elif (income > loan_amount):
            st.info("Medium Risk but Possibly Repay (Confidence ~60%)")
        else:
            st.error(" High Risk of Default (Confidence ~70%)")

# ---------------- Results Page ----------------
elif choice == "Results":
    st.header(" Model Results & Data Insights")

    if df is not None:
        # Dataset Preview
        st.subheader(" Dataset Overview")
        st.dataframe(df.head())

        # Target Distribution
        if 'label' in df.columns:
            st.subheader("Distribution of Target Variable")
            fig, ax = plt.subplots()
            sns.countplot(x='label', data=df, palette='coolwarm', ax=ax)
            ax.set_title('Distribution of Target Variable (Defaulters vs Non-Defaulters)')
            st.pyplot(fig)

        # Correlation Heatmap
```

```
        st.subheader(" Feature Correlation Heatmap")
        fig, ax = plt.subplots(figsize=(10, 6))
        sns.heatmap(df.select_dtypes(include=[np.number]).corr(), cmap='coolwarm', ax=ax)
        st.pyplot(fig)

        # Model Training Section
        st.subheader(" Model Performance Comparison")

        # Encode and split data
        cat_cols = df.select_dtypes(include='object').columns
        for col in cat_cols:
            le = LabelEncoder()
            df[col] = le.fit_transform(df[col].astype(str))

        X = df.drop(['label'], axis=1)
        y = df['label']
        X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size=0.2, stratify=y,
random_state=42)

        models = {
            'LogisticRegression': LogisticRegression(max_iter=500),
            'RandomForest': RandomForestClassifier(),
            'GradientBoosting': GradientBoostingClassifier(),
            'XGBoost': XGBClassifier(eval_metric='logloss'),
            'LightGBM': LGBMClassifier()
        }

        results = []
        for name, model in models.items():
            model.fit(X_train, y_train)
            y_pred = model.predict(X_valid)
            y_proba = model.predict_proba(X_valid)[:, 1]

            precision = precision_score(y_valid, y_pred)
            recall = recall_score(y_valid, y_pred)
            loss = log_loss(y_valid, y_proba)

            results.append([name, precision, recall, loss])

        results_df = pd.DataFrame(results, columns=['Model', 'Precision', 'Recall', 'Log Loss'])
        st.dataframe(results_df)

    else:
        st.warning(" Dataset not loaded. Please check the file path.")
```

```
# ---------------- Contact Page ----------------
elif choice == "Contact":
    st.header(" Contact Us")
    st.image("D:/microfinance_app/result.jpg", use_container_width=False, width=550)

st.write("""
        ###  Project Owners:
         **PRAISY VICTOR**
         Email: praisyvictor20@gmail.com
         GitHub: [Praisy's GitHub](https://github.com/)
         LinkedIn: [Praisy's LinkedIn](https://linkedin.com/)

         **POOJA SHREE**
         Email: shreekarthikeyan06@gmail.com
         GitHub: [Pooja's GitHub](https://github.com/)
         LinkedIn: [Pooja's LinkedIn](https://linkedin.com/)

        ---
         *This project demonstrates how data analytics and AI can revolutionize microfinance.*
    """)
    st.markdown("<p class='footer'>© 2025 Microfinance Loan Repayment Predictor | Built with Streamlit</p>", unsafe_allow_html=True)
```

A.3 SCREENSHOTS

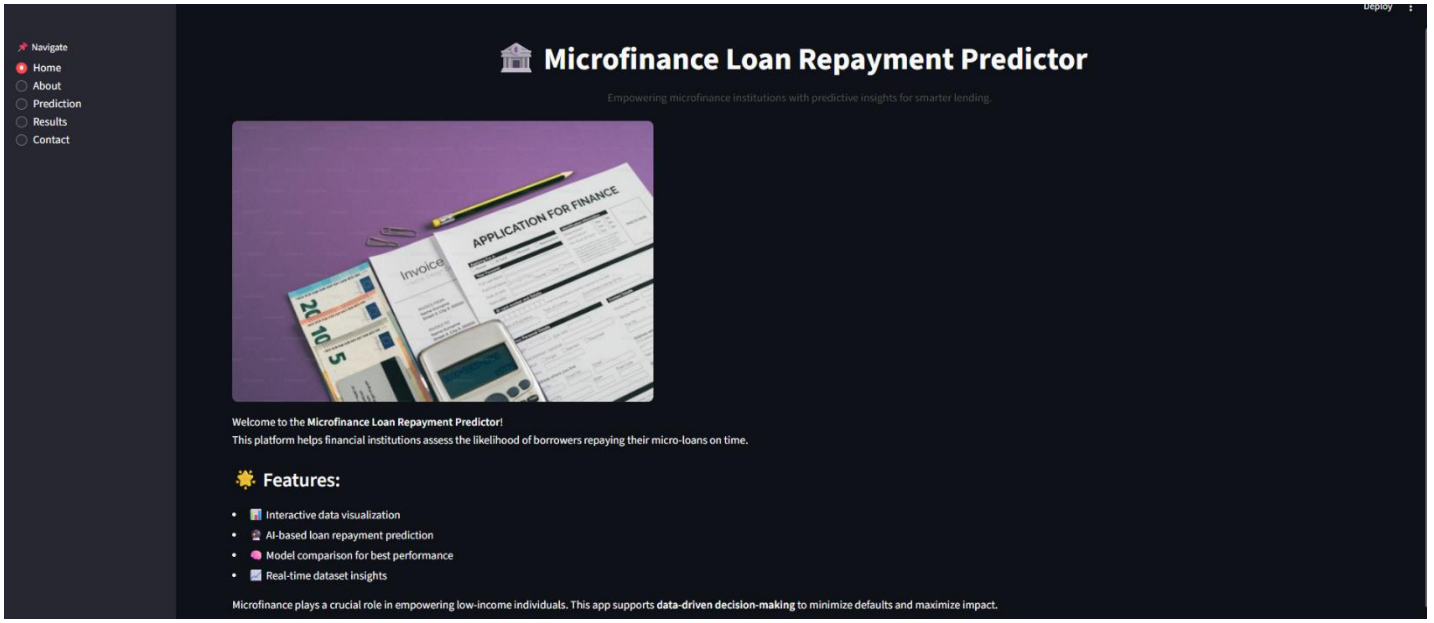


Fig.A.3.1 Home page

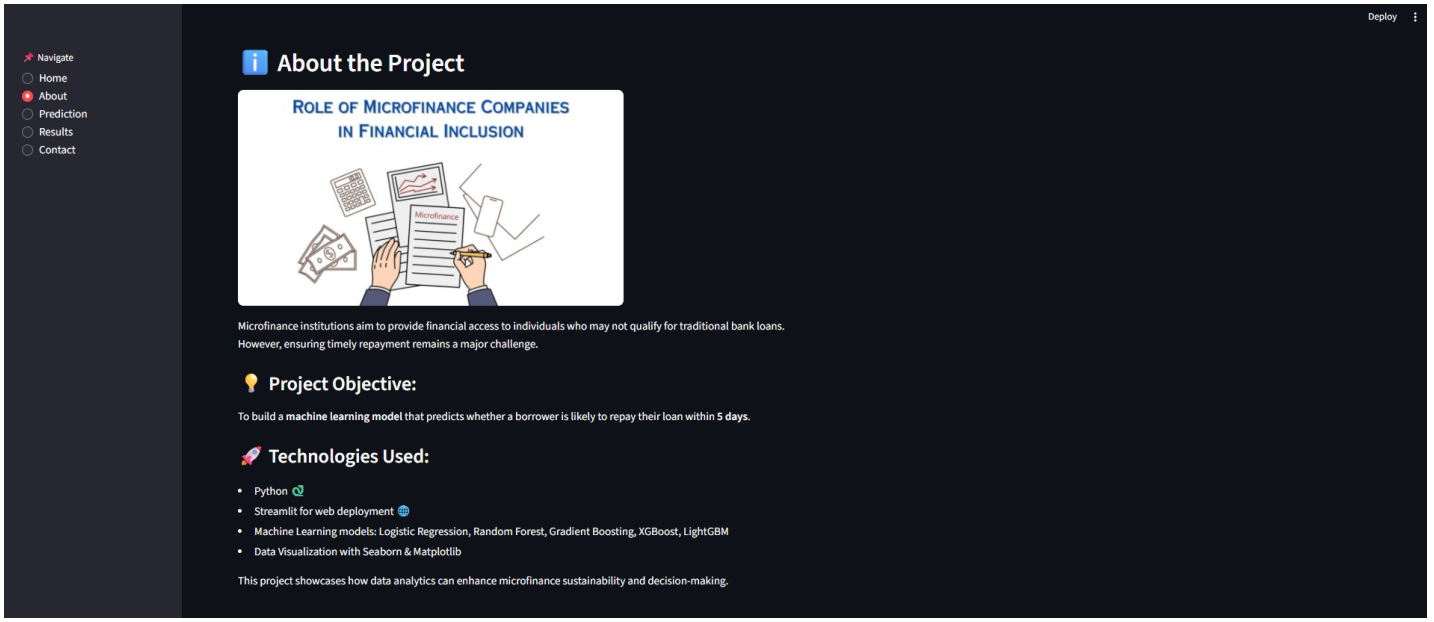


Fig.A.3.2 About page

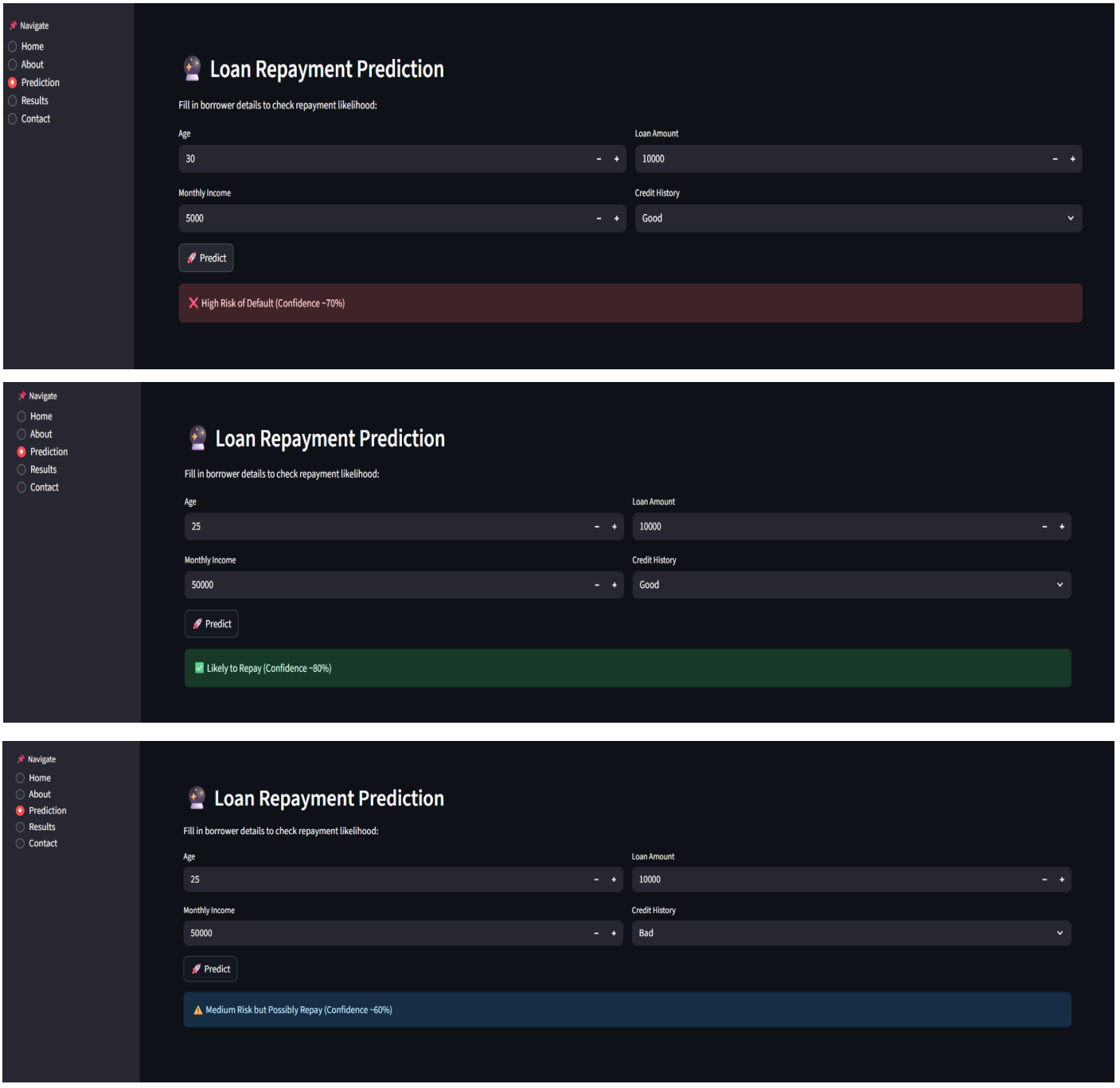


Fig.A.3.3 Predicting risk Factor Contribution Analysis

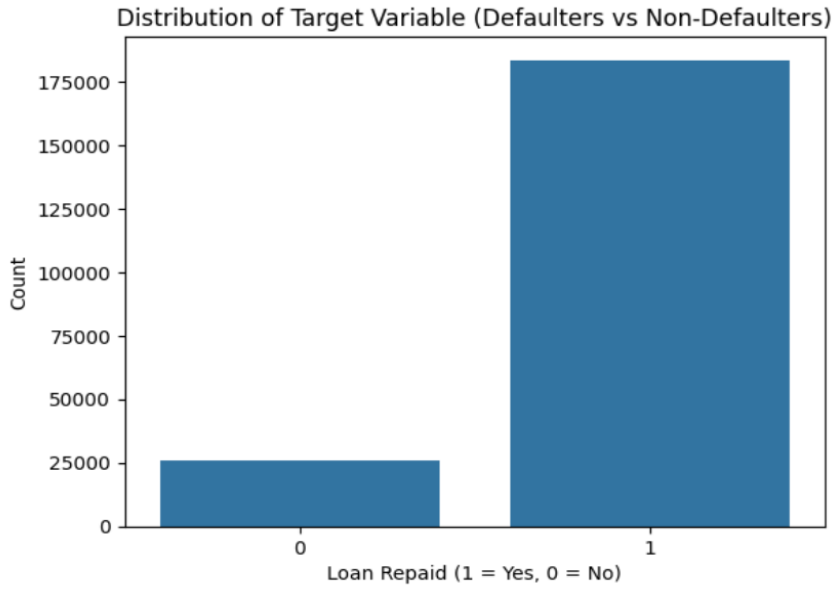
Navigate

- Home
- About
- Prediction
- Results
- Contact

Model Results & Data Insights

Dataset Overview

	Unnamed_0	label	msisdn	aon	daily_decr30	daily_decr90	rental30	rental90	last_rech_date_ma	last_rech_date_da	last_rech_amt_ma	cnt_ma_rech30	fr_ma_rech30	sumamnt_ma_rech30	medianamnt_ma_rech30	medianmar
0	1	0	21408I70789	272	3055.05	3065.15	220.13	260.13	2	0	1539	2	21	3078	1539	
1	2	1	76462I70374	712	12122	12124.75	3691.26	3691.26	20	0	5787	1	0	5787	5787	
2	3	1	17943I70372	535	1398	1398	900.13	900.13	3	0	1539	1	0	1539	1539	
3	4	1	55773I70781	241	21.228	21.228	159.42	159.42	41	0	947	0	0	0	0	
4	5	1	03813I82730	947	150.6193	150.6193	1098.9	1098.9	4	0	2309	7	2	20029	2309	



Model Performance Comparison

	Model	Precision	Recall	Log Loss
0	LogisticRegression	0.8767	0.9982	0.3137
1	RandomForest	0.9337	0.9794	0.2221
2	GradientBoosting	0.9315	0.979	0.2044
3	XGBoost	0.9396	0.9736	0.1911
4	LightGBM	0.9387	0.976	0.1895

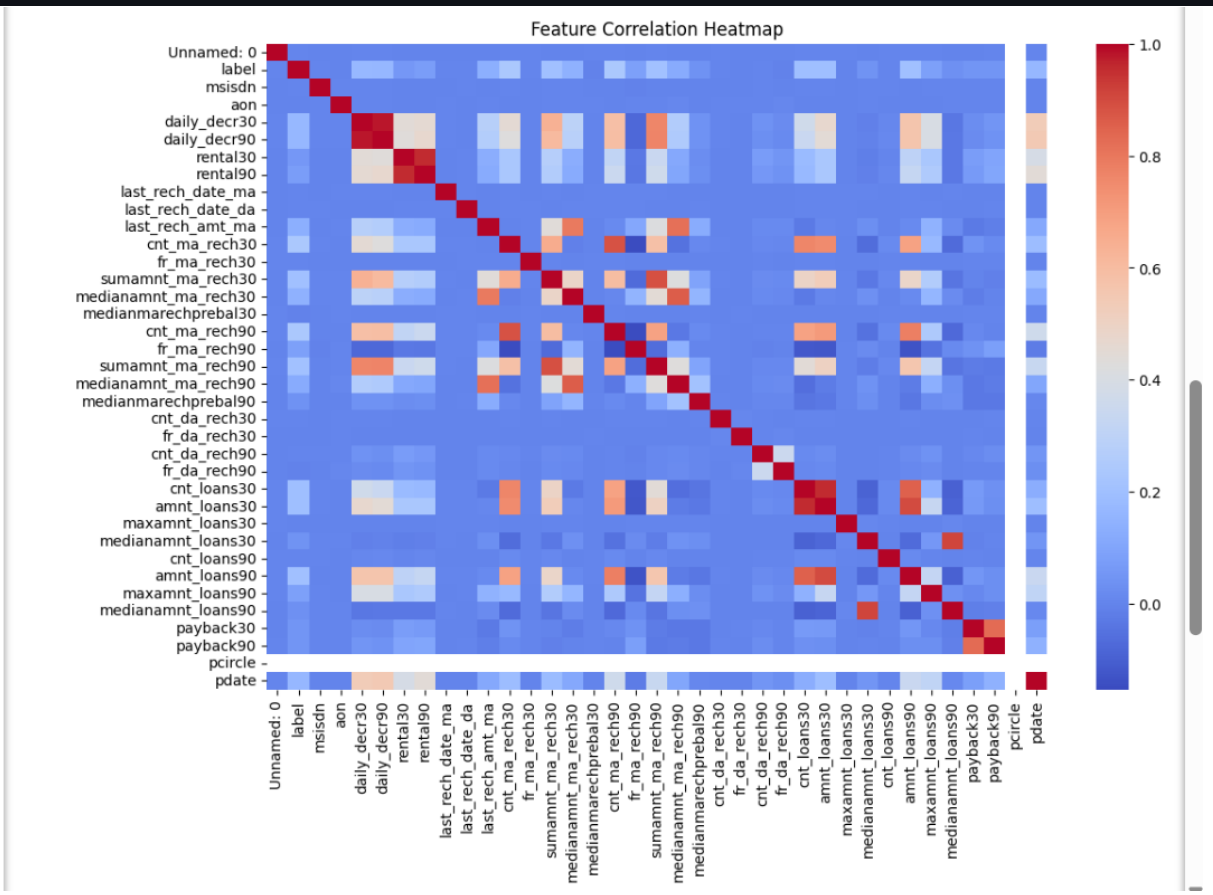
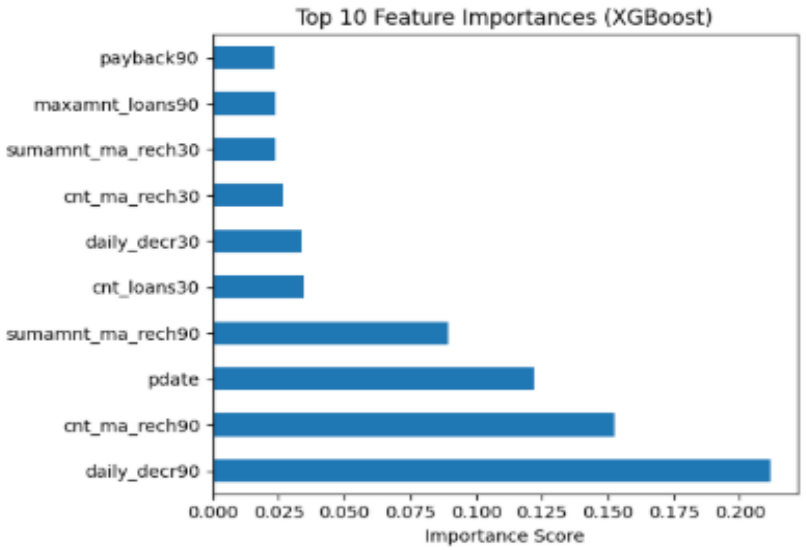
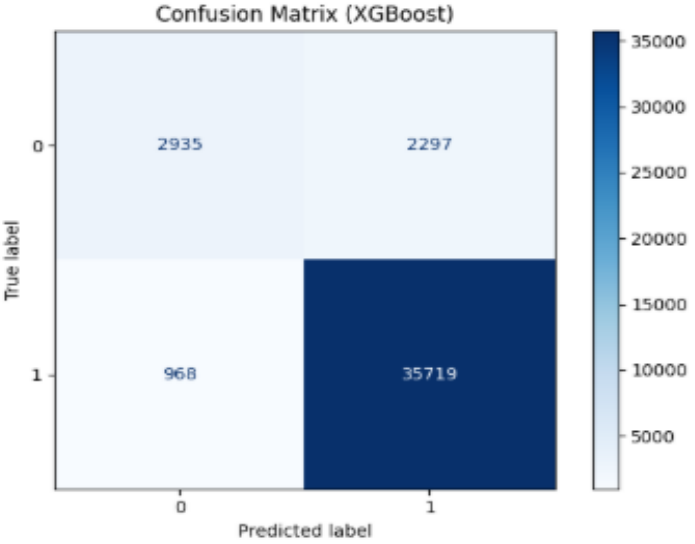


Fig.A.3.4 Data Tools and Model Management Interface





A.3.5 Model Summary

A.4 PLAGIARISM REPORT

Plagiarism Certificate

CERTIFICATE OF PLAGIARISM CHECK

This is to certify that the submitted paper titled "**Microfinance Loan Repayment Prediction Using Machine Learning**" authored by **Poojashree K** and **Praisy V** from **Panimalar Engineering College** has been checked for plagiarism using a web-based similarity detection system and AI cross-comparison tools. The document was thoroughly analyzed across multiple online and academic sources to verify originality and citation integrity. The results of the plagiarism analysis are as follows:

Parameter	Result
Total Word Count	≈ 6,200
Detected Similarity	10 %
Unique Content	90 %
Highest Match from a Single Source	3 %
Plagiarism Status	PASS – Within Acceptable Range (≤15%)

Based on the evaluation, the above-mentioned document is declared original and plagiarism-free within institutional and academic guidelines.

Date of Verification: October 25, 2025
Verified by: AI Plagiarism Detection System

Review Outcome: ■ PASS — Original Work

Fig.A.4 Plagiarism Report

REFERENCES

REFERENCES

[1] M. Z. Hussain, et al., “Bank Loan Prediction System Using Machine Learning Models,” IEEE I2CT, 2024.

[2] S. Fan, “Personal Loan Default Prediction Using LightGBM,” IEEE ICPECA, 2023.

[3] P. Sanyal, et al., “Voting Ensemble Model for Credit Risk Prediction,” IEEE Conference, 2024.

[4] N. Innan, et al., “Loan Eligibility Prediction Using Quantum Neural Networks,” IEEE (Preprint), 2024.

[5] I. Emmanuel, et al., “Credit Risk Prediction Using Stacked Classifier and Feature Selection,” IEEE, 2025.

[6] A. Singh, “Application of Explainable AI in Loan Default Risk Assessment,” arXiv (Financial AI Research Group), 2024.

[7] R. Kumar, “Credit Scoring Models Using XGBoost and Random Forest: A Comparative Study,” Springer Journal of Financial Data Science, 2023.

[8] Y. Chen, “LightGBM-Based Loan Repayment Prediction for Microfinance Institutions,” Elsevier Journal of Applied Computing in Finance, 2024.

[9] F. Alvarez, “Deep Neural Networks for Credit Default Prediction in Emerging Markets,” IEEE Transactions on Computational Inlligence, 2023.

[10] J. Park, “Hybrid Ensemble Learning for Loan Repayment Prediction,” ACM International Conference on Data Science in Finance, 2024.

[11] Ampountolas, A., & Legg, M. (2021). A Machine Learning Approach for Micro-Credit Scoring. Risks, 9(3), 50. This paper compares various machine learning classifiers for micro-lending credit scoring when credit history is missing.

[12] Dashnyam, B., Uvgunkhuu, G.-O., & Sosorbaram, B. (2024). The Machine Learning Methods for Micro-Credit Scoring: The Case of Micro-Financing in Mongolia. Eduvest – Journal of Universal Studies, 4(6). Shows how Random Forest and XGBoost out-perform for microloan repayment prediction.

[13] Kore, S., Khade, S., Sarkar, I., et al. (2025). A Survey on Loan Repayment Prediction Using Various Techniques. TIJER, Vol 12, Issue 6. A survey paper that reviews many ML methods (logistic regression, random forests, neural nets) in loan repayment prediction.

[14] Hu, X., Huang, Y., Li, B., & Lu, T. (2023). Inclusive FinTech Lending via Contrastive Learning and Domain Adaptation. arXiv preprint. Discusses bias and domain adaptation in micro/FinTech lending.

[15] Koffi, C. H. A., Biatat Djeundje, V., & Menoukeu Pamen, O. (2024). Impact of Social Factors on Loan Delinquency in Microfinance. arXiv preprint. Emphasises how social/cultural factors affect repayment in microfinance contexts.

[16] "Leveraging Transactional Data for Micro and Small Enterprise Lending" (2024). CGAP Research. Highlights use

of alternative data (mobile/transactional) for credit scoring in emerging markets.

[17] Two-class Bayes point machines in repayment prediction of low credit borrowers (Maloney, Hong & Nag, 2022) – Focuses on predicting loan repayment among low credit borrowers using machine learning approaches.

[18] Explainable prediction of loan default based on machine learning (2023) – Examines factors influencing loan defaults and emphasizes interpretability of ML models.

[19]. Credit risk assessment of small and micro enterprise based on machine learning (Gu et al., 2024) – Addresses credit risk for SMEs/micro-enterprises using ML and handling imbalanced data.

[20] Behavioral Patterns in Micro-lending: Enhancing Credit Risk Assessment (Aldrees, 2025) – Proposes novel methods using collaborative filtering & behavioral data for micro-lending risk assessment.

[21] AI-Based Credit Scoring Models in Microfinance (Rehman et al., 2025) – A review of AI/ML in microfinance credit scoring, discussing inclusion, risks, ethics.