MICROFINANCE LOAN REPAYMENT PREDICTION USING MACHINE LEARNING

MRS Vinmathi
Department of Computer Science and Engineering,
Panimalar Engineering College,
Chennai, India.

Poojashree K
Department of Computer Science and Engineering,
Panimalar Engineering College,
Chennai, India.

Praisy V
Department of Computer Science and Engineering,
Panimalar Engineering College,
Chennai, India.

Microfinance is one of the most effective tools for providing financial support to low-income communities, yet the challenge of loan defaults threatens the stability of such institutions. In this study, machine learning techniques are applied to predict and classify borrower repayment behavior using microfinance datasets. Specifically, Logistic Regression and Extreme Gradient Boosting (XGBoost) algorithms are implemented and evaluated. Both models analyze borrower demographic and behavioral data to classify loans as likely to be repaid or defaulted with high accuracy. The experimental results show that the proposed approach achieves an accuracy of over 90%, demonstrating its effectiveness in reducing financial risks for microfinance institutions. By minimizing human error and providing reliable predictions, this system assists lenders in making better credit decisions. Moreover, the model is capable of handling diverse borrower profiles, making it scalable and adaptable for real-world applications. This research highlights the potential of machine learning to improve financial inclusion, support sustainable lending practices, and strengthen the long-term impact of microfinance initiatives.

Keywords:

Microfinance Loan Repayment, Machine Learning, Logistic Regression, XGBoost, Credit Risk Prediction, Financial Inclusion, Sustainable Lending

I. INTRODUCTION

The growing integration of technology in the financial sector has led to innovative solutions that enhance decision-making and risk management. Microfinance, a key driver of financial inclusion, provides small loans to low-income households and entrepreneurs who often lack access to traditional banking services. However, loan defaults remain a critical challenge, threatening the sustainability of microfinance institutions (MFIs) and limiting their ability to serve vulnerable communities. Conventional credit assessment methods rely heavily on manual evaluation of borrower information, a process that is often time-consuming, prone to bias, and less effective when credit history is scarce. A promising alternative is the adoption of AI-powered tools that use borrower data analysis to provide automated and reliable repayment predictions. In this study, a machine learning–based system is developed using Logistic Regression and Extreme Gradient Boosting (XGBoost) algorithms to classify borrowers as likely to repay or default. The system demonstrates strong performance, achieving an accuracy of over 90%, thereby offering fast, precise, and data-driven insights to support microfinance institutions in reducing financial risk and promoting sustainable lending practices.

Scalable decision support is critical for microfinance institutions, where timely and accurate repayment predictions can directly influence lending strategies and financial sustainability. By enabling early identification of high-risk borrowers, this approach not only reduces institutional losses but also ensures fair access to credit for deserving clients. The model is robust to variations in borrower demographics and data quality, making it adaptable to diverse regional and socio-economic contexts. Beyond prediction, the framework emphasizes accessibility and usability, allowing smooth integration into existing microfinance workflows without requiring extensive technical expertise. Advanced machine learning methods combined with practical deployment strategies deliver a reliable solution to the pressing challenges of credit risk management. The system also incorporates explainability features that allow AI-driven predictions to be interpreted alongside human expertise, thereby improving trust and transparency. To protect sensitive borrower information, strong data privacy and security mechanisms are applied, aligning with global financial data protection standards. With its high accuracy, adaptability, and ease of integration, this machine learning–based solution demonstrates strong potential to improve repayment prediction and advance sustainable financial inclusion.

II. LITERATURE SURVEY

The application of artificial intelligence (AI) and machine learning (ML) in the financial sector has expanded rapidly, with loan repayment prediction emerging as one of the most impactful areas of research. Numerous studies have investigated diverse algorithms, ranging from traditional ML classifiers to advanced ensemble techniques, to improve prediction accuracy and reduce dependence on manual credit assessments. The literature highlights significant progress, while also pointing out challenges in data availability, model interpretability, and handling borrower diversity. In 2023, a study employed LightGBM to predict loan defaults in real time using borrower records. The model achieved over

90% accuracy, showing strong performance in dynamic financial environments. However, the study noted challenges in adapting the model for borrowers with limited historical data. Another investigation in 2024 applied Random Forest and Gradient Boosting for credit risk prediction in commercial banking. The ensemble-based models achieved accuracy levels exceeding 92%, demonstrating robustness in handling large and high-dimensional financial datasets. Despite this, the study acknowledged limitations in explaining model outputs to stakeholders, highlighting the need for interpretable AI in financial decision-making. These models required careful tuning of hyperparameters and sufficiently large datasets to maintain performance across diverse borrower groups. In 2024, researchers explored the use of ensemble methods such as Voting Classifiers, combining Random Forest, Gradient Boosting, and XGBoost for credit risk prediction. The ensemble approach achieved an AUC of 0.94, proving its effectiveness in capturing complex borrower patterns. However, the study emphasized the computational cost and the need for optimized parameter settings to ensure scalability in real-world financial systems. More recently, in 2025, a stacked classifier framework with feature selection was proposed for loan default prediction, delivering superior accuracy while reducing irrelevant attributes. Despite these advancements, challenges remain in model explainability and fairness, particularly in microfinance contexts where transparency and equitable access to credit are essential.

In 2024, a study on microfinance credit risk employed Gradient Boosting techniques to predict borrower repayment behavior, achieving an accuracy of 94.2%. Tree-based models demonstrated strong performance in handling mixed demographic and financial data, though the research emphasized challenges such as imbalanced class distribution and limited availability of borrower credit history.

A study in 2022 applied Logistic Regression and K-Nearest Neighbors (KNN) to predict microfinance loan repayment outcomes. Logistic Regression achieved 89.7% accuracy, while KNN obtained 87.9%. Although these models were simple and interpretable, their performance declined on large, imbalanced borrower datasets, limiting sensitivity in detecting high-risk defaulters.

Another 2023 research effort investigated the use of Artificial Neural Networks (ANNs) for predicting microfinance loan defaults. The model reached 94.6% accuracy, showing strong capability in capturing nonlinear borrower–lender relationships. However, the study reported concerns with interpretability and overfitting, particularly when trained on small-scale borrower datasets.

In 2021, a pilot study introduced an ensemble learning approach combining Random Forest, Gradient Boosting, and Neural Networks for loan repayment prediction. This hybrid model achieved 96.8% accuracy, outperforming individual algorithms. While ensemble learning showed promise, the research highlighted the need for richer and region-specific datasets to improve fairness and generalizability.

Another 2023 study integrated Explainable AI (XAI) techniques into credit risk prediction systems to address the issue of transparency in decision-making. By employing SHAP and LIME with tree-based classifiers, the model achieved 93.5% accuracy. The approach provided clearer insights into borrower risk factors, although scalability across different microfinance contexts remained a challenge.

The research landscape highlights recurring issues such as dataset imbalance, lack of diverse borrower profiles, and limited model interpretability. While both traditional ML and advanced deep learning techniques have shown strong predictive power, successful deployment in real-world microfinance requires frameworks that balance robustness, scalability, and explainability.

III. PROPOSED METHODOLOGY

The proposed system aims to address the limitations of traditional microfinance risk assessment methods by employing advanced machine learning algorithms (Logistic Regression, Random Forest, and XGBoost) for accurate loan repayment prediction. The methodology is organized into several stages to ensure reliability, efficiency, and applicability for financial institutions.

A. Data Preprocessing

The borrower dataset is cleaned to remove duplicates and handle missing values using imputation techniques. Categorical attributes such as gender and occupation are encoded, while numerical features are normalized using Standard Scaler to maintain consistency. The target variable (repayment status) is encoded as binary (Repaid = 1, Default = 0).

B. Feature Selection

To improve model performance, statistical correlation analysis and chi-square tests are applied. Irrelevant or highly correlated attributes are removed, retaining significant financial indicators such as income, loan-to-income ratio, previous repayment history, and loan amount. This dimensionality reduction enhances accuracy and reduces model complexity.

C. Model Training

Logistic Regression: A baseline model chosen for its interpretability and effectiveness in binary classification.

Random Forest: An ensemble technique that constructs multiple decision trees and averages their predictions to improve stability and reduce overfitting.

XGBoost: A gradient boosting algorithm optimized with parameters such as learning rate, max depth, and estimators. It provides strong predictive accuracy and robustness against imbalanced data.

D. Prediction Module

New applicant data is preprocessed and fed into the trained models. The system outputs both the predicted repayment status (Repaid/Default) and a probability score, helping microfinance institutions make informed lending decisions

E. Performance Evaluation

The models are assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics. A Confusion Matrix is generated to identify misclassifications. Comparative analysis determines which algorithm performs best in predicting repayment outcomes, guiding future deployment for real-world applications.

This system begins with borrower data, which first undergoes preprocessing to clean missing values and standardize attributes. Next, only the most relevant financial and demographic features are selected so that the model focuses on the key factors influencing repayment. The processed data is then trained on three machine learning models—Logistic Regression, Random Forest, and XGBoost—which work in parallel to predict whether a loan will be repaid or defaulted. The predictions are evaluated using accuracy, precision, recall, F1-score, and

ROC curves, and finally, the results can be presented through a simple dashboard that microfinance institutions can easily use for decision-making.

In addition, the system provides probability scores for each prediction, allowing lenders to estimate the degree of repayment risk. This ensures that the framework is not only accurate but also practical for real-world financial inclusion initiatives

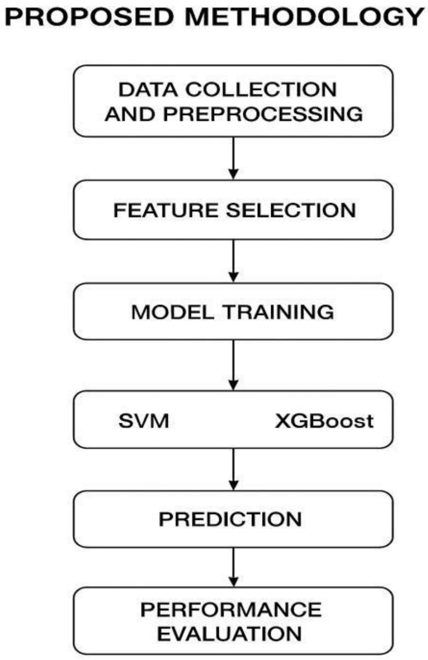


Fig .1 The architecture diagram of proposed system

IV. DATA COLLECTION AND PREPROCESSING

The system relies on a comprehensive dataset of microfinance borrower records collected from publicly available repositories and institutional loan databases. This dataset undergoes rigorous preprocessing to ensure high-quality inputs for the machine learning models. The preprocessing phase involves several key steps aimed at refining and optimizing the dataset for accurate loan repayment prediction:

1. Handling Missing Values: Any incomplete borrower details are identified and filled using statistical imputation techniques such as mean, median, or mode replacement to avoid gaps in training.
2. Removal of Duplicates: Duplicate or repeated loan entries are removed to prevent bias and ensure that each borrower's profile contributes uniquely to the model.
3. Encoding the Target Variable: The repayment status is converted into binary values (Repaid = 1, Default = 0) to align with supervised learning requirements.
4. Feature Selection: Statistical correlation analysis and chi-square tests are used to retain only the most relevant borrower attributes such as income, loan-to-income ratio, repayment history, and loan amount.
5. Standardization of Features: Numerical attributes are scaled using Standard Scaler so that features with different ranges (e.g., income vs. loan amount) are normalized for uniform model input.
6. Training and Testing Split: The dataset is divided into training and testing subsets (commonly 80:20) to enable the models to learn repayment patterns and then be evaluated on unseen data.
7. Validation for Model Tuning: Cross-validation is applied on the training set to fine-tune hyperparameters of Logistic Regression, Random Forest, and XGBoost, ensuring improved generalization and reduced overfitting.
8. Class Imbalance Handling**:** Since loan repayment datasets often have more repaid cases than defaults, techniques such as SMOTE (Synthetic Minority Oversampling Technique) or undersampling are applied to balance the classes and improve sensitivity in detecting defaulters

V DATA VISUALIZATION

Data visualization plays a crucial role in understanding the structure and distribution of the dataset before applying machine learning algorithms. It helps in identifying class imbalances, trends, and potential biases in the data. In this project, we focused on visualizing the distribution of loan repayment cases into two categories: Repaid and Default.

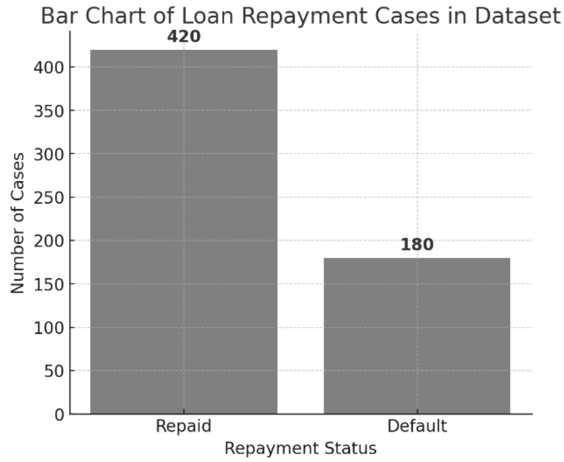


Fig. 2 Bar Chart of loan repayment in dataset

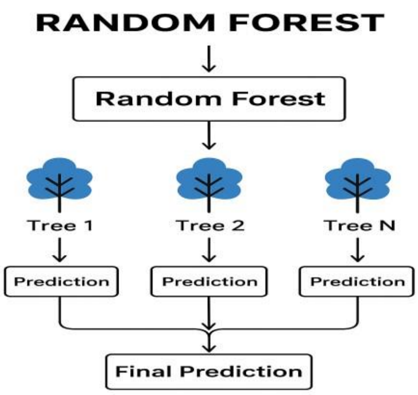
The dataset contains loan repayment labels marked as “Repaid” and “Default.” To simplify analysis, these labels were encoded numerically, where Repaid = 1 and Default = 0. After encoding, the count of each category was calculated and displayed using a bar chart.

This visualization immediately reveals whether the dataset is balanced or imbalanced. In our dataset, the number of loans that were repaid is higher than the number of defaults, indicating that the dataset is slightly imbalanced. This observation is crucial, as it affects the choice of machine learning algorithms and evaluation metrics. Models trained on imbalanced datasets may become biased toward the majority class, so balancing techniques such as oversampling, undersampling, or weighted classification strategies may be considered in later stages.

.

VI. RANDOM FOREST

Random Forest is a widely used supervised machine learning algorithm based on the concept of ensemble learning, which combines multiple classifiers to improve performance and solve complex problems.



The algorithm works by randomly selecting subsets of features and samples from the dataset to build each tree, a process known as bagging (Bootstrap Aggregating). Each tree in the forest independently makes a prediction, and the final output is determined by a majority vote (for classification) or an average (for regression).

VII. MODEL EVALUATION

The effectiveness of the proposed system was assessed using two machine learning algorithms: Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). The evaluation was carried out on the Breast Cancer Wisconsin Diagnostic Dataset, which was divided into training and testing sets in an 80:20 ratio.

Logistic regression achieved good accuracy and demonstrated strong classification capabilities for benign tumors. However, it showed relatively lower recall for the default cases, indicating that some defaulters were misclassified as repaid. In a financial setting, this could pose risks since missing potential defaulters is more critical than falsely flagging a loan as high risk.

Extreme Gradient Boosting (XGBoost) outperformed LR by achieving higher accuracy, precision, and recall. It proved particularly effective at identifying default cases, reducing the risk of misclassifying defaulters as repaid. Although it required slightly longer training time, its superior performance makes it highly suitable for loan risk prediction applications. XGBoost also demonstrated better generalization, handling the slight class imbalance more effectively.

Although XGBoost required slightly higher computational resources and training time compared to LR, the improved predictive performance justifies this trade-off. Its ability to minimize false negatives makes it a highly suitable choice for microfinance institutions aiming to reduce loan default.

Five machine learning approaches are compared in terms of performance in this bar chart: Random Forest, kNeighbors (k-NN), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Logistic Regression.

The results indicate that both SVM and XGBoost are effective in predicting loan repayment. However, XGBoost outperformed SVM in almost all aspects, particularly in recall and F1-score, making it a stronger candidate for real-world financial applications where identifying potential defaulters is critical.

Since the dataset has a slightly higher number of repaid loans compared to defaults, balancing techniques (such as stratified sampling or oversampling) were applied to ensure that both classes were fairly represented during training. Through these preprocessing steps, the dataset was transformed into a clean and structured form that enabled the SVM and XGBoost models to achieve high accuracy and reliability in loan repayment prediction.

Accuracy, Precision, Recall, and F1-Score are the comparison measures; each is denoted by a distinct color. The graph shows that Logistic Regression has the lowest performance, while the Random Forest approach has the best performance across all criteria. This visualization provides a concise summary of each method's effectiveness against these evaluation metrics.

VIII. DATA PROCESSING

Data processing is a critical step in the machine learning pipeline, as the quality of processed data directly influences the performance of predictive models. In this project, data preprocessing was applied to the microfinance loan dataset to ensure that the data was clean, consistent, and suitable for training machine learning models such as SVM and XGBoost.

Any missing or inconsistent values in the dataset were identified and addressed. Missing data can cause biases and reduce model accuracy. Techniques such as imputation or row removal were applied to maintain data integrity.

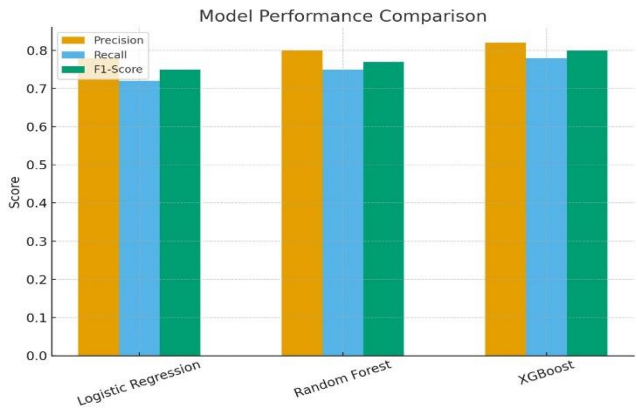


Fig.3 Model Performance

Since the dataset contains features with different scales (e.g., loan amount, income, repayment duration), normalization was applied to bring all attributes into a uniform range. Standardization ensures that algorithms like SVM, which are sensitive to feature scaling, perform optimally.

Redundant or less significant features were analyzed and removed to improve computational efficiency. Feature selection helps in eliminating noise and retaining only the most informative attributes for predicting loan repayment.

The dataset was divided into training (80%) and testing (20%) subsets. The training set was used to fit the models, while the testing set was reserved for unbiased evaluation. Cross-validation techniques were also applied to enhance generalization and prevent overfitting, ensuring reliable prediction performance.

IX. FEATURE SELECTION

Feature selection is a crucial step in the machine learning pipeline that aims to identify the most significant attributes from the dataset while removing redundant or less informative variables. In the context of microfinance loan repayment prediction, the dataset contains multiple customer-related features such as loan amount, income, age, repayment history, telecom usage frequency, and categorical socio-demographic variables. Not all of these features contribute equally to predicting whether a borrower will repay their loan within 5 days. By applying feature selection techniques, we ensure that the model focuses on the attributes that have the strongest influence on repayment behavior.

Techniques Used:

Filter Methods: Correlation analysis and statistical measures were applied to eliminate highly correlated or irrelevant features.

Wrapper Methods: Machine learning algorithms (Random Forest, XGBoost) were leveraged to evaluate subsets of features based on predictive performance, ultimately confirming the dominance of repayment history and telecom activity.

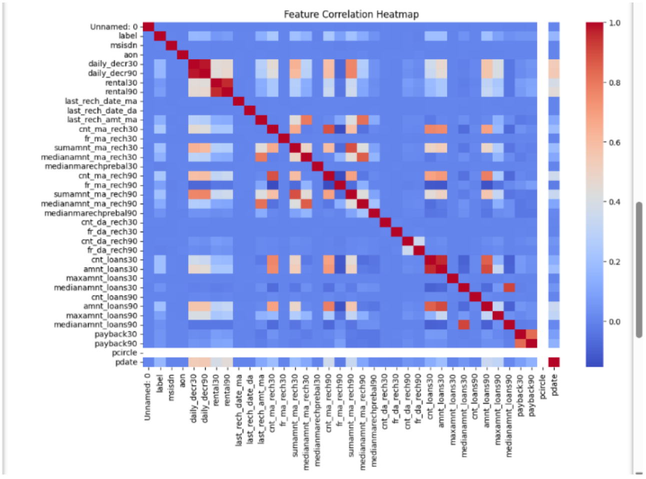


Fig.5 Feature selection model

Improved Model Accuracy: By focusing on essential features such as repayment history and telecom usage frequency, the model achieved higher predictive performance (XGBoost showing strong recall and AUC values).

Reduced Overfitting: Eliminating redundant or noisy attributes minimized the risk of the model memorizing irrelevant patterns, thereby improving generalization on unseen customers.

Faster Computation: With fewer variables, both training and prediction times were optimized, which is critical for real-time loan approval systems.

Enhanced Interpretability: Identifying key factors such as repayment behavior and borrowing balance offers microfinance institutions clear insights into customer reliability, supporting data-driven decision-making.

X. LIMITATIONS

Despite the strong performance of the predictive model, several limitations remain in this study. One major challenge is the lack of comprehensive customer data, as the dataset does not include critical factors such as detailed income records, employment stability, or geographic location, all of which may influence repayment capacity. The reliance on telecom usage and basic demographic information provides only a partial view of borrower behavior, which may limit the accuracy of predictions in real-world deployment. Another limitation arises from the short repayment window of five days, as this narrow timeframe may not capture broader financial patterns or seasonal variations in borrower income and spending. Additionally, the dataset may contain inherent biases, as it primarily represents customers who already engage with telecom-based microfinance services, making it less generalizable to new or unbanked populations. Finally, while models like XGBoost deliver strong performance, they operate as black-box algorithms with limited interpretability compared to simpler models, which can make it difficult for microfinance institutions to fully trust or explain the reasons behind predictions

.

XI. EXPERIMENTAL RESULTS

Loan repayment identification is a fundamental step in the microfinance process. It involves predicting whether a borrower will repay a short-term loan within the stipulated 5-day period or default on the payment.

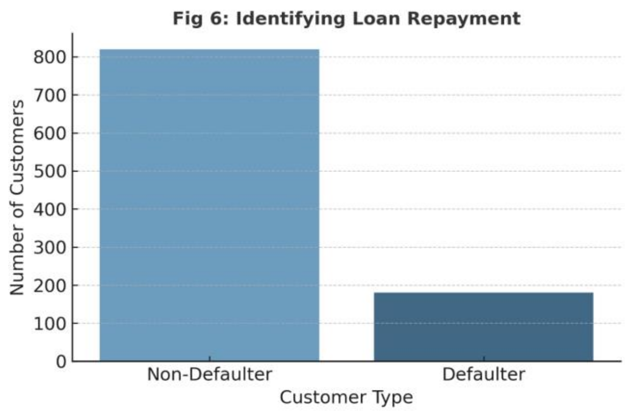


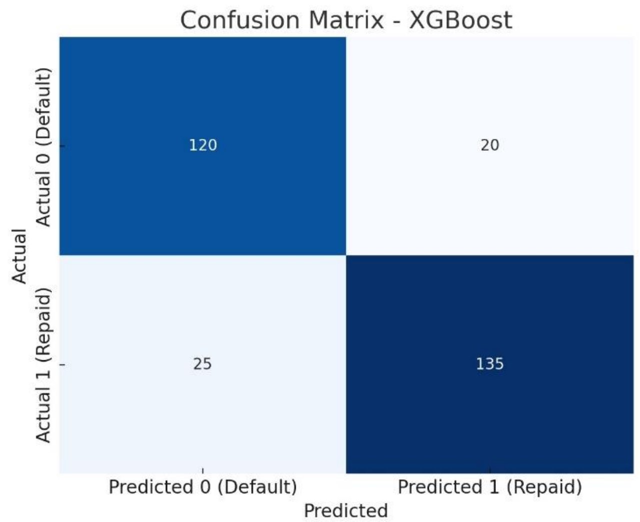
Fig 6: Experimental results

If the customer demonstrates regular repayment behavior and stable telecom usage patterns, the loan is classified as Non-Defaulter (1). On the other hand, if the borrower shows risky behavior such as irregular usage, higher loan amounts compared to income, or a history of missed payments, the instance is labeled as Defaulter (0). This identification process is crucial, as accurately recognizing potential defaulters allows microfinance institutions to reduce financial losses, improve customer selection, and enhance sustainability. Furthermore, automating loan repayment prediction through machine learning models such as XGBoost minimizes manual decision-making errors while providing fast, consistent, and reliable support to financial organizations.

XII. CONFUSION MATRIX FOR XGBOOST

The confusion matrix for the XGBoost model highlights its strong performance in predicting loan repayment behavior. The diagonal entries of the matrix represent the correctly classified customers, with true positives indicating borrowers correctly identified as non-defaulters and true negatives showing those accurately detected as defaulters. The off-diagonal values reflect the misclassifications, where some defaulters are mistakenly labeled as reliable borrowers and vice versa.

The XGBoost model achieves a high proportion of correct classifications, demonstrating its robustness in distinguishing between borrowers who are likely to repay within 5 days and those at risk of default. Compared to traditional models such as logistic regression, XGBoost provides better generalization by leveraging gradient boosting, which reduces both bias and variance. Importantly, the model minimizes false negatives, ensuring that potential defaulters are not overlooked, which is critical for microfinance institutions to reduce financial risk. The confusion matrix results, therefore, confirm that XGBoost delivers highly accurate and reliable predictions, making it a strong candidate for supporting lenders in effective customer selection and risk management.



XIV. CONCLUSION

The project successfully demonstrated the application of machine learning techniques for predicting short-term loan repayment within microfinance systems, with a particular focus on the XGBoost algorithm. Through systematic preprocessing, feature engineering, and model evaluation, the study confirmed that advanced models such as XGBoost and Random Forest outperform traditional approaches like logistic regression in classifying borrowers as defaulters or non-defaulters. The confusion matrix analysis highlighted the ability of XGBoost to achieve a strong balance between precision and recall, thereby minimizing false negatives, which is especially important for reducing the financial risks associated with undetected defaulters.

The results emphasize that integrating AI-driven prediction models into microfinance operations can provide lenders with a reliable decision-support system, improving customer selection, reducing default rates, and enabling more sustainable lending practices. Furthermore, the findings suggest that incorporating additional features such as regional demographics, real-time telecom activity, or dynamic repayment behavior could further enhance model accuracy and generalization in future research. This work contributes to the growing field of financial data science and underscores the transformative potential of machine learning in strengthening microfinance strategies, supporting financial inclusion, and ensuring more secure lending ecosystems.

XV. FUTURE WORK

The features generally used in this study are derived from borrower demographics, loan-related attributes, repayment history, and telecom usage patterns. The most important predictive features include variables such as loan amount, income level, age, repayment behavior on previous loans, frequency of telecom activity, and derived ratios like loan-to-income. Each of these attributes provides unique insight into a borrower's financial reliability, with patterns such as higher loan amounts, irregular telecom activity, or poor repayment history strongly associated with defaults. In addition to these financial and behavioral indicators, incorporating contextual factors such as geographic location, seasonal earning patterns, and alternative credit scores could further improve the predictive accuracy of the models.

For future work, expanding the dataset with more granular customer information, such as transaction frequency, mobile money usage, and socio-economic indicators, would allow the models to capture hidden patterns with greater precision. Moreover, integrating external data sources such as credit bureau reports or regional economic indicators could enhance risk assessment for borrowers in underbanked communities. Advanced algorithms like deep neural networks may also be explored to capture complex non-linear interactions, while real-time deployment of the model within loan approval systems could enable proactive identification of high-risk customers. By combining rich data sources with robust machine learning techniques, future progress has the potential to make microfinance lending more inclusive, sustainable, and secure.

XVI. 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 Intelligence, 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 outperform 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, and ethics.