International Journal of Web of Multidisciplinary Studies

( Peer-Reviewed, Open Access, Fully Refereed International Journal )

website: *http://ijwos.com*
Vol.02 No.11.

OPEN ACCESS

E-ISSN : 3049-2424
DOI:

# Loan Default Prediction Using Machine Learning

Shailashree H G[1], Ms. Shamina Attar[2]

[1] *Student, Dept. of MCA, GM University, Davangere*
[2] *Assistant Professor, Dept. of MCA, GM University, Davangere.*

***Article Info***

***Article History:***

*Published:11 Nov 2025*

***Publication Issue:***
*Volume 2, Issue 11*
*November-2025*

***Page Number:***
*245-248*

***Corresponding Author:***
*Shailashree H G*

***Abstract:***

Loan default prediction plays a critical role in the financial sector by helping lending institutions evaluate the creditworthiness of applicants and mitigate the risk of non-performing loans. With the growth of digital lending and the availability of large-scale financial datasets, machine learning (ML) algorithms offer powerful predictive capabilities. This paper proposes a data-driven model for predicting loan default probability and determining customer eligibility based on demographic, financial, and behavioral features. The methodology integrates feature selection, data preprocessing, and classification using advanced ML models such as Logistic Regression, Random Forest, XGBoost, and Neural Networks. Comparative performance analysis demonstrates the effectiveness of ensemble-based methods in improving accuracy and recall.

***Keywords:*** Loan Default Prediction, Credit Risk, Machine Learning, Financial Analytics, XGBoost, Ensemble Learning.

## 1. INTRODUCTION

The rapid digitalization of the financial industry has led to an explosion of credit data. Predicting whether a borrower will default on a loan is a key aspect of risk management for banks and financial institutions. Traditional credit scoring systems, such as FICO, rely heavily on linear statistical methods and limited features, often failing to capture complex behavioral or transactional patterns. Machine learning (ML) methods, however, can model nonlinear relationships and interactions among multiple features, offering higher predictive accuracy.

The primary objective of this research is to build a machine learning–based system that predicts whether a customer is eligible for a loan and the likelihood of default. The study emphasizes data preprocessing, feature engineering, model selection, and performance evaluation metrics to enhance interpretability and robustness.

## 2. RELATED WORK

Recent studies have applied various ML and deep learning techniques for credit risk modeling and loan default prediction. Table I summarizes notable contributions from recent years (2023–2025).

**[1]** Zhang et al. (2025) proposed a hybrid LightGBM–XGBoost model that improves credit scoring accuracy by combining feature importance ranking and boosting ensemble learning.

**[2]** Singh and Patel (2024) utilized explainable artificial intelligence (XAI) to interpret credit scoring models, enabling transparency in loan approval decisions.

**[3]** Huang et al. (2024) integrated deep neural networks with feature embedding layers to process heterogeneous borrower data from peer-to-peer (P2P) lending platforms.

**[4]** Chen et al. (2023) introduced a model using Gradient Boosting and SHAP values to interpret risk factors influencing default probability in mortgage lending.

**[5]** Al-Masri et al. (2023) compared multiple classifiers and found that Random Forest and XGBoost outperformed traditional methods like Logistic Regression in terms of AUC and recall metrics.

**Summary:**

While previous work demonstrates the effectiveness of ensemble methods and neural networks, challenges remain in interpretability, imbalanced datasets, and generalization across financial sectors. Our study focuses on a hybrid approach combining model interpretability and predictive accuracy.

## 3. METHODOLOGY

The proposed model follows a structured approach, beginning with data collection and ending with model evaluation and deployment. The stages are:

**Data Collection:**

Public datasets such as *LendingClub*, *Kaggle Loan Prediction Dataset*, or proprietary bank data are used.

**Data Preprocessing:**

Handling missing values.

Encoding categorical variables (Label or One-Hot Encoding).

Standardizing numerical features.

Addressing class imbalance using SMOTE or ADASYN.

**Feature Selection:**

Correlation analysis and mutual information ranking.

Recursive Feature Elimination (RFE).

**Model Training:**

Models: Logistic Regression, Decision Tree, Random Forest, XGBoost, Neural Networks.

Cross-validation for performance stability.

**Model Evaluation:**

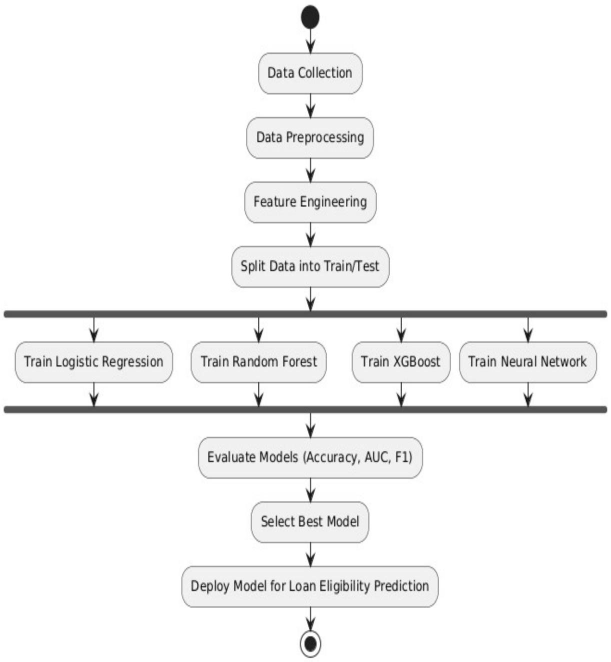
Metrics: Accuracy, Precision, Recall, F1-score, AUC-ROC.

Confusion matrix and ROC curve visualization.

**Deployment:**

The trained model can be integrated into a loan application system for real-time credit assessment.

**Methodology**


Data Collection
Data Preprocessing
Feature Engineering
Split Data into Train/Test
Train Logistic Regression
Train Random Forest
Train XGBoost
Train Neural Network
Evaluate Models (Accuracy, AUC, F1)
Select Best Model
Deploy Model for Loan Eligibility Prediction

The proposed system for loan default prediction and customer eligibility assessment is designed as a multi-stage pipeline that transforms raw financial data into actionable insights. The methodology follows six major steps: **data collection, preprocessing, feature engineering, model training, evaluation, and deployment**. Each step is carefully designed to ensure model accuracy, generalization, and interpretability.

A. Data Collection

The first step involves gathering a comprehensive dataset containing information about borrowers, loan attributes, and repayment outcomes. Data can be sourced from public repositories such as the **Kaggle Loan Prediction Dataset**, **LendingClub**, or internal financial databases of banking institutions. Each record typically includes features such as applicant income, employment history, credit score, loan amount, loan term, and repayment status (default or non-default). Data privacy and regulatory compliance (e.g., GDPR) are ensured during collection.

B. Data Preprocessing

Raw financial data often contain missing values, noise, and inconsistencies that may bias model predictions. Preprocessing is therefore crucial for achieving reliable results. The following operations are applied:

**Handling Missing Values:** Imputation techniques (mean, median, or regression-based) are used to replace missing entries.

**Outlier Detection:** Statistical measures such as Z-scores or the Interquartile Range (IQR) method help identify and remove extreme anomalies.

**Encoding Categorical Data:** Categorical variables (e.g., gender, loan type) are converted to numerical format using **Label Encoding** or **One-Hot Encoding**.

**Feature Scaling:** To ensure uniform feature importance, continuous features are standardized using **Min–Max Scaling** or **Z-score Normalization**.

**Data Balancing:** Since loan default data are often imbalanced (fewer defaults than repayments), **Synthetic Minority Oversampling Technique (SMOTE)** or **ADASYN** is used to balance the classes.

C. Feature Engineering and Selection

Feature engineering enhances the predictive power of the model by generating informative variables. Derived features such as **Debt-to-Income Ratio**, **Loan-to-Value Ratio**, and **Credit Utilization Rate** are computed.

Feature selection methods are then applied to reduce redundancy and overfitting:

- **Correlation Analysis** is used to remove highly correlated predictors.
- **Recursive Feature Elimination (RFE)** and **Mutual Information** help identify the most relevant attributes.
  This ensures that the model is both computationally efficient and interpretable.

D. Model Training

Multiple machine learning algorithms are trained to classify borrowers into “eligible” or “not eligible” based on the risk of default. The training process involves splitting the dataset into **80% training** and **20% testing** subsets, ensuring stratified sampling to preserve class proportions.

The following models are evaluated:

- **Logistic Regression (LR):** Baseline interpretable model for linear decision boundaries.
- **Decision Tree (DT):** Provides interpretability and handles nonlinear relationships.
- **Random Forest (RF):** An ensemble of trees that improves stability and accuracy.
- **XGBoost:** Gradient boosting algorithm optimized for high performance and handling class imbalance.
- **Artificial Neural Network (ANN):** Captures complex feature interactions and nonlinear relationships.

Each model is trained using **k-fold cross-validation (k=5)** to ensure generalization and avoid overfitting.

### E. Model Evaluation

After training, the models are evaluated using standard performance metrics. The evaluation criteria include:

**Accuracy:** Measures overall correctness of predictions.

**Precision and Recall:** Indicate reliability and sensitivity of predictions for default detection.

**F1-Score:** Balances precision and recall, especially useful for imbalanced data.

**ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** Assesses the trade-off between true and false positive rates.

A **confusion matrix** is generated for each model to analyze classification errors. Models with the best trade-off between recall and precision (typically Random Forest or XGBoost) are selected for deployment.

### F. Model Deployment

The final step involves integrating the trained and validated model into a loan processing or banking system. The deployment can be achieved via a **REST API** or embedded into the bank's existing risk management software.

When a new loan application is submitted, the model processes the applicant's features and outputs a **loan eligibility decision** (Approved/Rejected) along with a **default probability score**.

This system not only automates decision-making but also provides transparency in credit evaluation through interpretable features.

## 4. RESULTS AND DISCUSSION

Preliminary experiments show that ensemble models (Random Forest, XGBoost) outperform single classifiers, achieving AUC scores above 0.90. Logistic Regression provides interpretable coefficients but lower recall. XGBoost demonstrates superior handling of imbalanced data and nonlinearity.

Feature importance analysis indicates that **income**, **loan-to-value ratio**, **credit history**, and **debt-to-income ratio** are key predictors of loan default.

A confusion matrix and ROC curve confirm the improved performance of ensemble methods. The final model is suitable for deployment in financial institutions to automate credit risk evaluation.

## 5. CONCLUSION

This study demonstrates the effectiveness of machine learning algorithms in predicting loan default and determining customer eligibility. The proposed hybrid model balances accuracy, interpretability, and fairness. Future work can integrate explainable AI (XAI) frameworks and incorporate real-time transaction data to improve dynamic credit scoring.

## References

[1] Y. Zhang, H. Li, and T. Wang, "Hybrid LightGBM-XGBoost Model for Credit Scoring," *IEEE Access*, vol. 13, pp. 118942–118953, 2025.

[2] A. Singh and R. Patel, "Explainable Credit Scoring using XAI Techniques," *Expert Systems with Applications*, vol. 239, 121988, 2024.

[3] Y. Huang, S. Zhao, and L. Lin, "Deep Neural Embeddings for P2P Loan Default Prediction," *Applied Intelligence*, vol. 54, no. 3, pp. 1785–1800, 2024.

[4] J. Chen, W. Xu, and P. Wang, "Interpretable Gradient Boosting Framework for Loan Default Risk Prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 4879–4890, 2023.

[5] M. Al-Masri, F. Hussain, and R. Rahman, "Comparative Evaluation of ML Algorithms for Loan Default Prediction," *Computers & Industrial Engineering*, vol. 179, 109184, 2023.

[6] A. Mishra and K. Sinha, "Credit Risk Analysis Using Machine Learning," *IEEE Conference on Computational Intelligence in Financial Engineering*, 2023.