*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 3, Ser. 3 (May. – June. 2025), PP 01-04*
*www.iosrjournals.org*

# "Loan Default Prediction"

Mahina Mazed, Misbah Maheen, Misba Kounain, Kushalini K R,
Dr. Omar Khan Durrani

***Abstract***
*Traditionally, the manual workload and time involved in the loan approval process is intensive. In this paper, an automated approach is provided by the authors through the creation of a Loan Prediction Model, which leverages machine learning to predict the outcome of loan approvals using applicant information. The proposed model classifies the loan approval status of an applicant using Logistic Regression, XGBoost, and Decision Trees. In addition, a hybrid model that aggregates the outcomes of the separate models is evaluated. With the Kaggled dataset which included income, credit history, and employment details of applicants, the models developed were able to predict outcomes. While the outcome accuracy by the models was not satisfactory, the analysis performed laid groundwork concerning applying multiple machine learning systems to assist in loans prediction, and highlighted key obstacles related to model performance sufficiency.*

Date of Submission: 17-05-2025 Date of Acceptance: 27-05-2025

## I. Introduction

Loan approval is perhaps the most important and frequent decision-making process that banks and financial institutions make. Historically, the process has been done manually by credit officers and underwriters who evaluate applicants based on a plethora of factors including income, employment history, credit scores, and current liabilities. While the human judgment superiority of manual evaluation exists, the procedure is lengthy, subject to inconsistencies, vulnerable to bias, and subject to human error. As there is a growing demand for quick financial services as well as accuracy, there exists a need to automate loan approval with the help of technologies.

The scope of this research is to make use of ML algorithms in order to automate and enhance the loan approval process. Through training the predictive models using past loan information, we hope to be able to make accurate predictions on whether an application for a loan will most likely be granted. The models used in this research are Logistic Regression, XGBoost, and Decision Trees—each renowned for their ability to perform classification tasks and detect hidden patterns in data. Additionally, we discuss the feasibility of applying these models in a hybrid system to enhance overall predictive quality and system reliability.

We aim to create an objective, efficient, and data-driven loan prediction model with low bias, low human intervention, and increased operational efficiency. Major variables examined during analysis include the income level of the applicant, credit history, employment status, and other pertinent financial parameters. Through the use of ML for loan approval automation, banks can decrease the risk of fraudulent approvals, make the process transparent and fair, and offer quicker service to customers. The research overall intends to help lead to a wiser, more trustworthy, and more just financial system through the integration of machine learning into basic banking processes.

## II. Litrature Survey

Machine learning (ML) prediction of loan default is now an essential area of study in the banking industry due to the significant monetary risks involved for defaulting borrowers. Recent research has investigated a range of ML methods, from basic statistical models to sophisticated ensemble techniques, in order to enhance predictive performance and lending decision-making processes.

The initial research by Shaheen and ElFakharany is centered on comparative evaluation of different ML algorithms, such as statistical (logistic regression, decision trees), AI-based (neural networks), and ensemble methods (Random Forest and Gradient Boosted Trees) for loan default prediction based on a large dataset from an Egyptian government-owned bank. Their results showed that ensemble approaches performed better than single classifiers, with Random Forest giving 91.7% accuracy and 95.83% precision, and Gradient Boosted Trees providing 91.9% accuracy and 79.68% precision. This indicates that aggregating multiple learners dramatically improves prediction performance in credit risk modeling.[6]

In this paper presented at the 2020 IEEE CCNS conference, emphasizes the use of boosting techniques, specifically XGBoost and AdaBoost, in the context of consumer and peer-to-peer (P2P) lending.

The authors describe a robust feature extraction process from user attributes, lending history, and credit reports, followed by model evaluation using metrics such as accuracy and AUC. They point out that the boosting algorithms, with their capacity to handle intricate data patterns and mitigate overfitting, are best suited for noisy and imbalanced datasets commonly encountered in financial applications.[7]

In this paper, Ouyang offers a direct comparison between logistic regression and XGBoost in loan default prediction, especially in the context of rural financial assistance in China. With cleaned and feature-engineered bank loan data, the research proves that XGBoost outperforms logistic regression in accuracy and AUC score. In addition, it uses SHAP value analysis to explain model outputs and determine important predictive features like income, debt-to-income ratio, and credit history. The research highlights the importance of advanced ensemble techniques for interpretable and efficient risk assessment, particularly in newer fields such as agricultural finance.[5]

Together, they highlight the shift towards more advanced ensemble and boosting approaches in credit risk modeling due to their better prediction ability, flexibility in handling sophisticated data sets, and capacity to generate understandable insights.

## III. Methodology

The data utilized in this research is obtained from Kaggle, describing a sequence of characteristics including applicant income, credit score, loan amount, and employment. The process can be described as the following steps:

**Data Preparation:**

The original dataset was put through a cleaning and preprocessing step to handle missing values, outliers, and categorical variables. The dataset was standardized to ensure that all the features were within a standard range, a necessary step for some machine learning algorithms like Logistic Regression and XGBoost. After the preprocessing step, features like "Loan_Status" were encoded, thus preparing the dataset for model training.

**Model Training:**

The information was divided into training and test sets (training 80% and testing 20%). We then trained three machine learning models:
logistic regression: A simple linear classifier as the default model.
XGBoost: A strong boosting algorithm known for its superior performance in classification tasks.
Decision Trees: A non-linear model with interpretability.

Also, we tested the viability of taking a hybrid model approach that takes an average of the predictions made by these different models. The idea was that using the strength of different models would improve overall prediction accuracy.

**Model Evaluation:**

The models were tested by the use of accuracy, precision, recall, and F1-score. These measurement metrics allow for an insight into the model's performance, especially in the case of imbalanced datasets. Although accuracy was the main measure, we also considered precision and recall to measure the models' performance in handling false positives and false negatives.

**Hybrid Model Implementation:**

The hybrid model was developed through the combination of the predictions generated by Logistic Regression, XGBoost, and Decision Trees. Ensemble techniques, such as voting classifiers and stacking, were attempted utilizing the goal of improving the total predictive performance.

**Architecture of the Model**

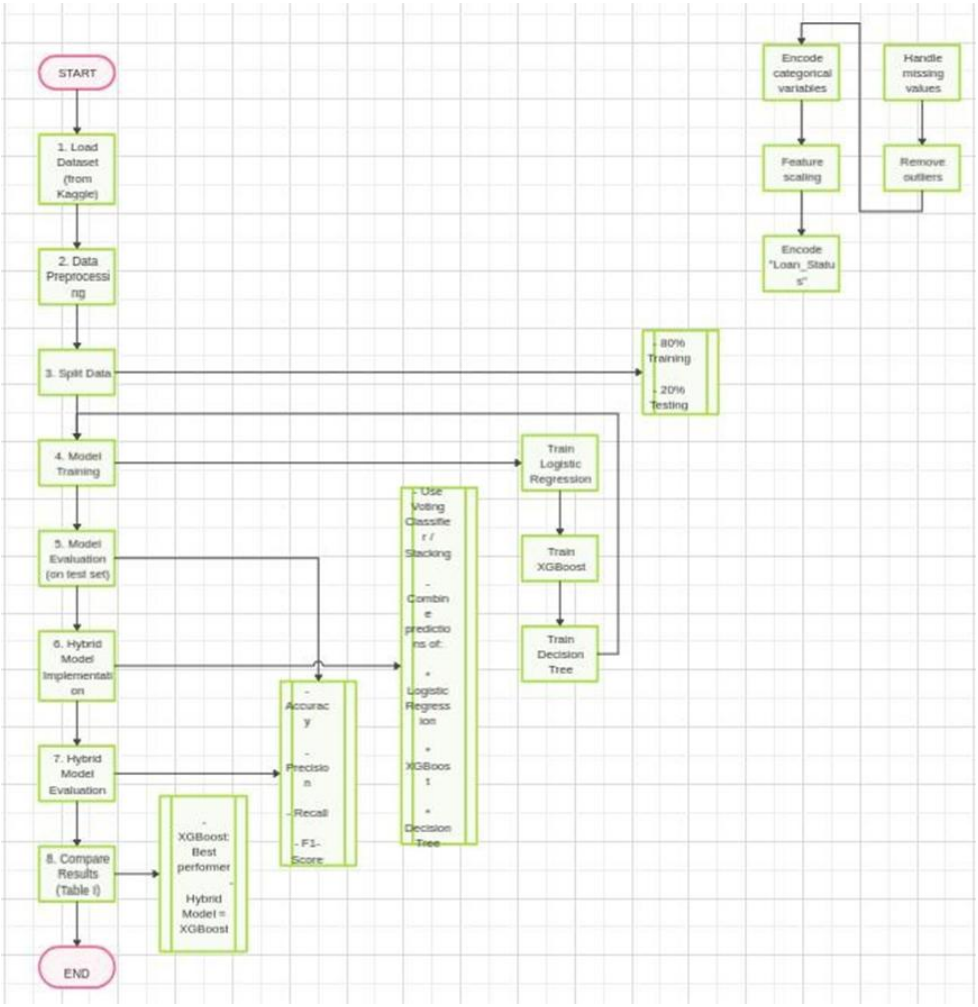


Fig: Model Architecture

The Model architecture in your study adheres to a machine learning ensemble pipeline.

Below is a systematic outline of the architecture and components:
Data Input and Pre-processing Pipeline
Input: Raw data from Kaggle (features such as applicant income, credit score, loan amount, employment, and target "Loan_Status").

Preprocessing steps:
Deal with missing values Identify and treat outliers.
Encode categorical variables(e.g., one-hot or label encoding).
Feature scaling (standardization) to normalize numerical features.

Model Training Architecture Data Split:
- 80% for training
- 20% for testing Models Trained:
- Logistic Regression (Linear classifier)
- XGBoost (Ensemble boosting algorithm)
- Decision Tree Classifier (Non-linear, interpretable model)

Hybrid Model (Ensemble Layer)
Once the three models were trained, their predictions were combined with ensemble methods:
Voting Classifier (presumably soft voting or hard voting)
Possibly stacking, where base model predictions are inputs to a meta-learner (although stacking wasn't discussed in your overview).

Evaluation Layer Metrics Used:
- Accuracy (main one)
- Precision
- Recall
- F1-score

These measures aid in measuring performance, particularly on unbalanced datasets.

## IV. Observation And Result

The performance of both the individual models and the hybrid combined model is shown in Table I. Surprisingly, the hybrid model did not perform significantly better than the individual models, suggesting that the synergies between the models were not achieved. The XGBoost model achieved the highest accuracy among the individual models, with an accuracy rate of 82%, followed by Logistic Regression at 78%. However, the performance of the hybrid model was on par with that of XGBoost, indicating that the use of the individual models independently may be more suitable for the problem in question.

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic Regression | 78 | 0.74 | 0.71 | 0.72 |
| XGBoost | 82 | 0.80 | 0.75 | 0.77 |
| Decision Tree | 75 | 0.72 | 0.68 | 0.70 |
| Hybrid Model | 82 | 0.79 | 0.74 | 0.76 |

**Table I: Evaluation of Model Effectiveness**

The XGBoost model performed better when it came to classification accuracy, indicating that boosting techniques may be better than the simpler models like Logistic Regression. The Decision Tree, however, even with interpretability, performed worse, especially in precision and recall. The Hybrid Model did not contribute much to enhancing the performance of XGBoost, which indicates that the models employed collectively did not provide a synergistic effect to the ultimate prediction. This result supports the need to carefully choose and optimize models in the process of building hybrid systems.

## V. Conclusion

This study proposes a Loan Prediction Model with machine learning models such as Logistic Regression, XGBoost, and Decision Trees. While each model performed well individually, the combined model did not provide a significant increase in the accuracy of prediction. The results show that the chosen models do not have a synergistic interaction, i.e., each model might perform individually better. Nevertheless, this study demonstrates the ability of machine learning to make the loan approval process automatic and thus reduce the errors of human intervention and improve operational efficiency. Future work in research can include the study of other ensemble algorithms such as Random Forest and Gradient Boosting Machines (GBM), and study of different hyper-parameter tuning for further improving the model's performance. Additionally, more features such as the applicant's past loan history or other external data sources can be added to improve the precision of prediction.

## Reference

[1] Aurélien Géron. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition. O'Reilly Media.
[2] VanderPlas, J. Python Data Science Handbook. O'Reilly Media.
[3] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill Education.
[4] Domingos, P. (2012). "A Few Useful Things to Know About Machine Learning." Communications of the ACM.
[5] "Loan Default Prediction Based on Logistic Regression and XGBoost Modeling" by Yi Ouyang.
[6] "Predictive analytics for loan default in Banking sector using Machine Learning technique" by Shaheen and ElFakharany.
[7] "Loan Default Prediction with Machine Learning Techniques" by 2020 International Conference on computer Communication and Network Security.