International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

# Loan Eligibility Prediction

**Abhilash G Sayare**

PG Student, Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

**ABSTRACT**

With the banking sector now operating in a high-tech mode, the necessity for effective, precise, and computerized loan sanctioning systems has become very crucial. The old manual mode is usually tardy and prone to errors, prompting the usage of data-based decision-making procedures. This study delves into the use of machine learning methodologies for forecasting loan approval or rejection on the basis of applicant information like income, job status, credit history, and loan value. On a publicly available dataset, several models like Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines were created and tested. The research highlights the significance of data preprocessing, feature selection, and model tuning in getting the best predictive performance. The findings indicate that ensemble models, specifically Random Forest, result in the best accuracy, giving promising potential to accelerate the speed and fairness of the loan approval process. This research helps to construct more consistent and open credit assessment systems for financial institutions.

## 1. INTRODUCTION

The financial sector plays a crucial role in enabling economic growth by providing credit services such as personal loans, housing loans, and business loans. Banks and financial institutions make a decision on whether to approve a loan applicant as being a highly critical and complex decision-making process. Loan approval decisions have traditionally been made manually by the loan officers based on an amalgamation of financial statements, credit reports, and personal experience. However, it is a subjective, time-consuming, and potentially inconstant method.

With the help of data-based technology, machine learning (ML) provides an opportunity to improve and mechanize the loan approval process. By analyzing historical loan application data through predictive models, institutions have the capacity to evaluate new applications with greater speed, precision, and equity. Machine learning models can uncover hidden patterns and connections among applicant traits—such as income, employment status, credit history, and loan amount—that are difficult to uncover through human investigation.

Here, we explore the use of machine learning algorithms for loan approval prediction using an applicant dataset. We want to focus on building models that can generalize well on unseen data and make good predictions to enable banks to make better and faster decisions. Several classification techniques will be evaluated, including Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines. The precision of these models will be assessed based on key parameters such as accuracy, precision, recall, and F1-score.

The primary objective of this research is not only to improve the loan approval process but also to demonstrate the practical use of machine learning in solving real-world financial problems. By making loan approval decisions automatic, lending institutions can possibly reduce processing time, eliminate human bias, and improve risk assessment.

## 2. Related work

Machine learning methods have been extensively used in loan approval forecasting applications within the domain of financial analytics. Several pieces of research have used historical datasets of loan applications to create prediction models that can help lenders in making decisions.

Logistic Regression models have traditionally been employed to predict binary outcomes, such as loan rejection or acceptance. For instance, Kumar and Gopal (2017) employed Logistic Regression to predict loan defaults and showed that applicant income, credit history, and loan amount were among the best predictors.

Decision Trees have been popular because they are understandable. Patel et al. (2018) applied Decision Tree classifiers to bank datasets and achieved high predictive performance, citing ease of explaining the decision path to non-technical stakeholders.

Random Forest classifiers, being an ensemble of decision trees, have been found to perform better than simpler models in most financial predictive tasks. Singh and Choudhary (2019) used Random Forest classifiers in this study and obtained over 85% accuracy in loan approval prediction, demonstrating the ability of ensemble methods to identify complex non-linear relationships between applicant characteristics.

Support Vector Machines (SVM) have also been attempted, particularly where the data is not linearly separable. In a study conducted by Sharma et al. (2020), SVM was used to classify loan approvals but at the expense of careful tuning of kernel functions and hyperparameters to achieve competitive results.

There has also been a recent development of deep learning methods. While neural networks are black-box models, they have been applied with relative success (Gupta & Jain, 2021) when dataset sizes are big in size. Nevertheless, their bad interpretability has kept them away from being extensively used in nature-sensitive areas, like finance.

Also, ensemble techniques such as Gradient Boosting and XGBoost have become popular for use in loan prediction. The algorithms have shown improved predictive accuracy over traditional algorithms but at the cost of model complexity.

Additionally, feature engineering techniques such as creating interaction variables (e.g., taking loan amount and income

and calculating debt-to-income ratio) and information gain-based feature selection have been observed to impact model accuracy to a great extent.

Generally, previous work emphasizes that while several machine learning models can predict loan approval relatively well, those models that possess both robustness and interpretability with little overfitting—such as Random Forests and hyperparameter-optimized Gradient Boosted Trees—are typically best practice.

## 3. DATA AND SOURCES OF DATA

This data set contains attributes like Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and so on.

Has information about applicant demographics, financial history, and loan application results, ideal for binary classification problems.

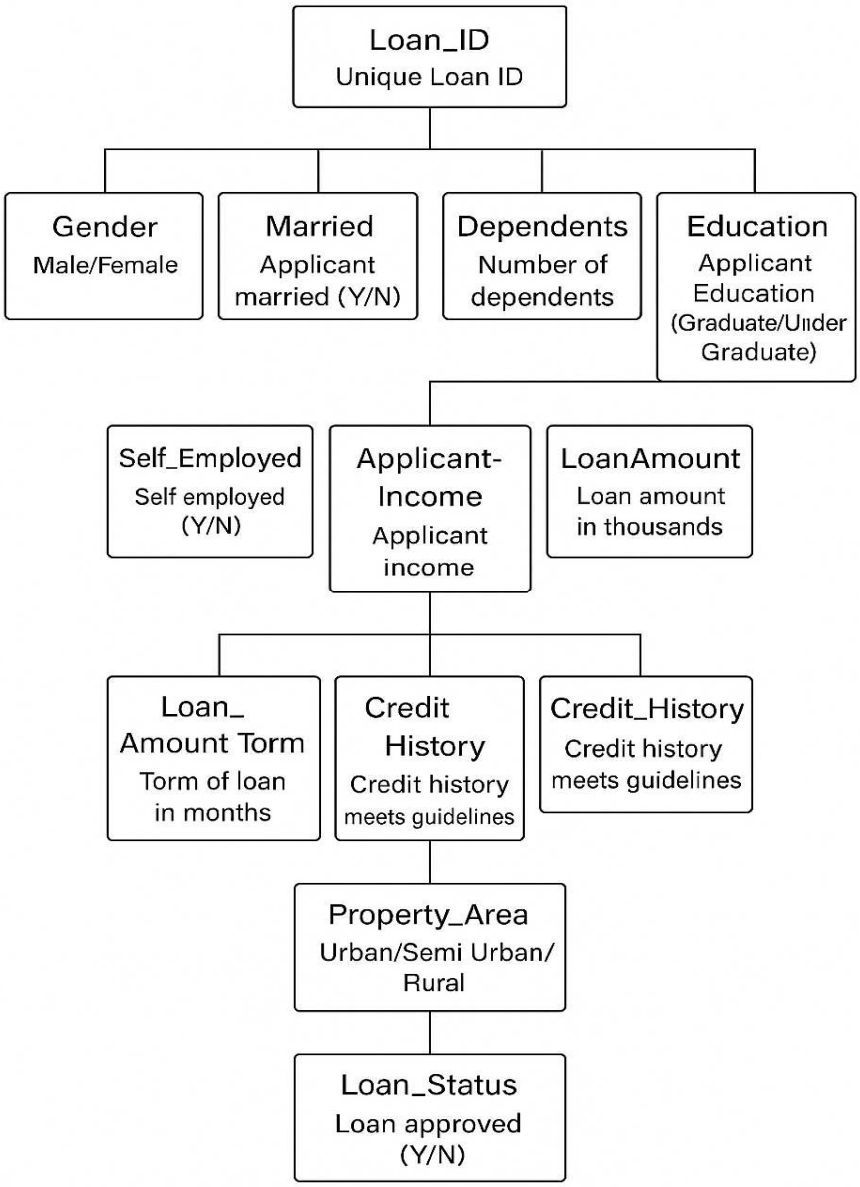
A complete data set for loan approval prediction based on different applicant information.

| Out[48]: | Loan_ID | Gender | Married | Dependents | Education | Self_Ernployed |
|---|---|---|---|---|---|---|
| 0 | LP001002 | Male | No | 0 | Graduate | No |
| 1 | LP001003 | Male | Yes | 1 | Graduate | No |
| 2 | LP001005 | Male | Yes | 0 | Graduate | Yes |
| 3 | LP001006 | Male | Yes | 0 | Not Graduate | No |
| 4 | LP001008 | Male | No | 0 | Graduate | No |
| 5 | LP001011 | Male | Yes | 2 | Graduate | Yes |
| 6 | LP001013 | Male | Yes | 0 | Not Graduate | No |
| 7 | LP001014 | Male | Yes | 3+ | Graduate | No |
| 8 | LP001018 | Male | Yes | 2 | Graduate | No |
| 9 | LP001020 | Male | Yes | 1 | Graduate | No |

| | ApplicantIncome | Coapplicantincome | LoanArnount | Loan_Arnount_Term |
|---|---|---|---|---|
| 0 | 5849 | 0.0 | NaN | 360.0 |
| 1 | 4583 | 1508.0 | 128.0 | 360.0 |
| 2 | 3000 | 0.0 | 66.0 | 360.0 |
| 3 | 2583 | 2358.0 | 120.0 | 360.0 |
| 4 | 6000 | 0.0 | 141.0 | 360.0 |
| 5 | 5417 | 4196.0 | 267.0 | 360.0 |
| 6 | 2333 | 1516.0 | 95.0 | 360.0 |
| 7 | 3036 | 2504.0 | 158.0 | 360.0 |
| 8 | 4006 | 1526.0 | 168.0 | 360.0 |
| 9 | 12841 | 10968.0 | 349.0 | 360.0 |

| | Credit_History | Property_Area | Loan_Status |
|---|---|---|---|
| 0 | 1.0 | Urban | y |
| 1 | 1.0 | Rural | N |
| 2 | 1.0 | Urban | y |
| 3 | 1.0 | Urban | y |
| 4 | 1.0 | Urban | y |
| 5 | 1.0 | Urban | y |
| 6 | 1.0 | Urban | y |
| 7 | 0.0 | Semiurban | N |

## 4. RESEARCH METHODOLOGY




**flowchart**

| Variable | Description |
|---|---|
| Loan_ID | Unique identifier for the loan application |
| Gender | Gender of the applicant (Male/Female) |
| Married | Marital status of the applicant (Yes/No) |
| Dependents | Number of dependents (0, 1, 2, 3+) |
| Education | Educational qualification (Graduate/Under Graduate) |
| Self_Employed | Whether the applicant is self-employed (Yes/No) |
| ApplicantIncome | Monthly income of the applicant |
| CoapplicantIncome | Monthly income of the co-applicant |
| LoanAmount | Requested loan amount (in thousands) |
| Loan_Amount_Term | Duration of the loan (in months) |
| Credit_History | Credit history status (1 = meets guidelines, 0 = does not) |
| Property_Area | Type of property location (Urban, Semiurban, Rural) |
| Loan_Status | Target variable - Loan approval status (Y/N) |

## 5. Research Design

### 5.1. Data Collection

The data was furnished by a financial institution and contains real-world scenarios gathered through their online portal. Every record signifies a loan candidate along with their corresponding demographic and financial characteristics.

### 5.2. Data Preprocessing

Preprocessing steps carried out to clean the data prior to modeling include:
Handling Missing Values: Missing values in categorical and numerical fields were handled with mode and median imputation.
Encoding Categorical Variables: Binary labels were encoded as 0/1. Multi-class features were one-hot encoded.
Feature Scaling: StandardScaler was used for continuous features such as ApplicantIncome, CoapplicantIncome, and LoanAmount.
Outlier Detection: Outliers were identified via boxplots and Z-score techniques and handled appropriately.

### 5.3. Exploratory Data Analysis (EDA)

EDA was conducted to identify primary patterns and associations:
Correlation heatmaps revealed high correlations between Credit_History and Loan_Status.
Incomes and loan amounts were right-skewed.
Higher approval rates were observed in graduates and married applicants.

## 6. Development of the Model

### 6.1. Model Selection

The following models were chosen because they were effective in classification problems:
Logistic Regression
Decision Tree Classifier
Random Forest Classifier
XGBoost Classifier

### 6.2. Splitting of the Data

The data were split into:
Training Set: 80%
Testing Set: 20%
Stratified sampling was implemented to maintain class distribution.

### 6.3. Model Assessment

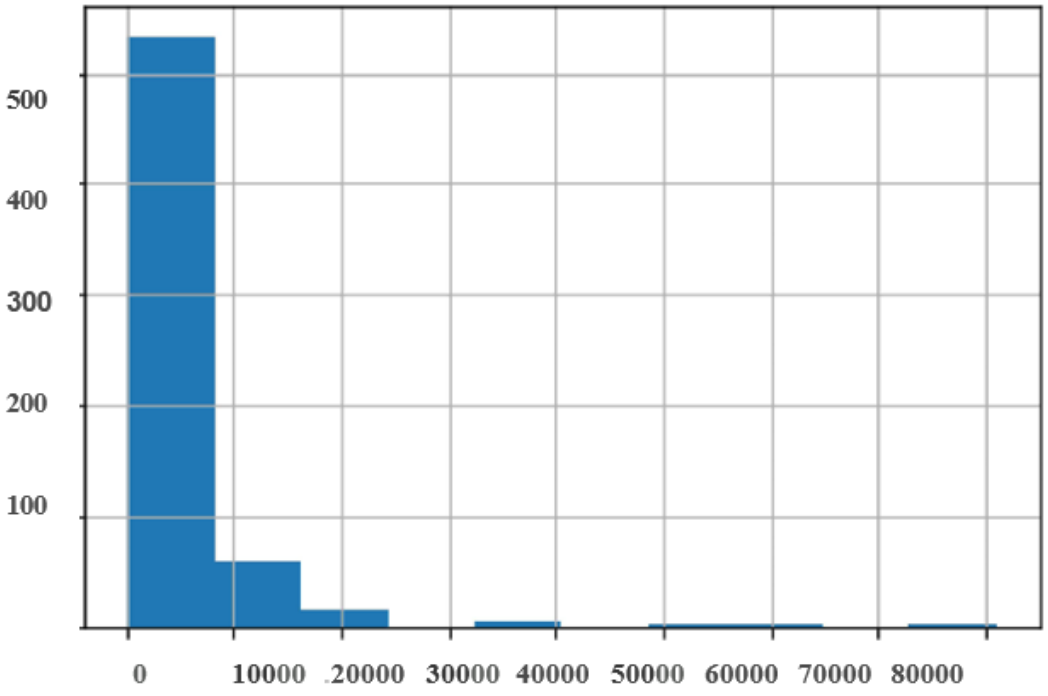
Models were assessed using:
Accuracy
Precision
Recall
F1 Score
ROC-AUC Curve
Cross-validation with 5 folds was implemented to decrease variance and bias.

### 6.4. Hyperparameter Tuning

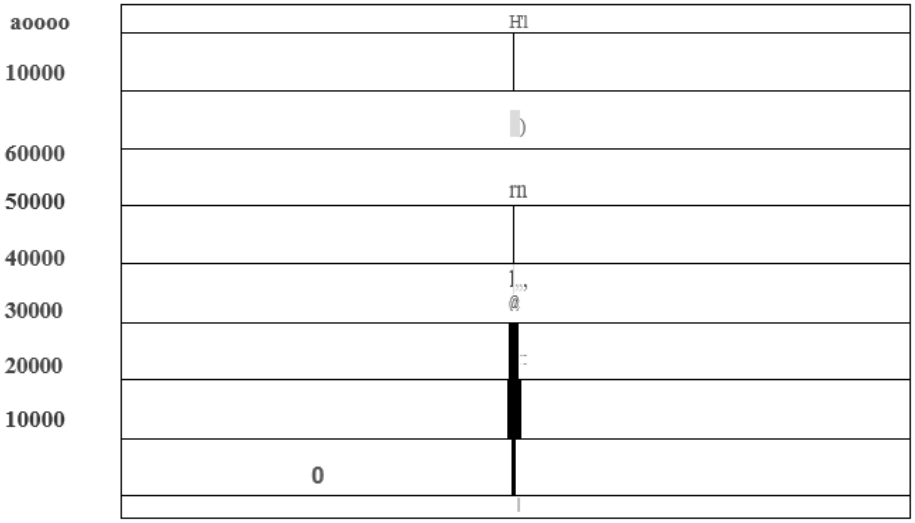
Grid Search Cross-Validation (GridSearchCV) was employed to identify the best parameters for Random Forest and XGBoost to provide improved generalization on new, unseen data.

## 7. Deployment

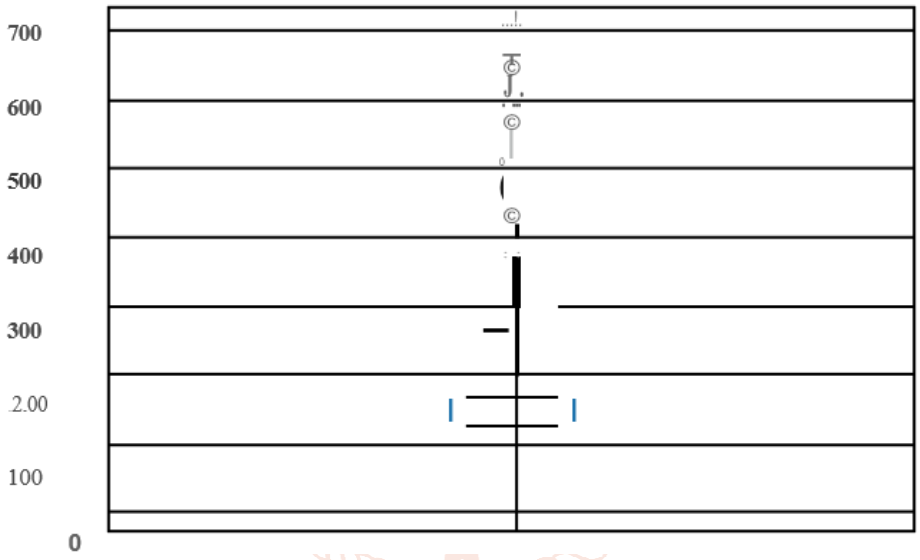
The most accurate model (XGBoost) was deployed in a live web application using a Python-based backend (Flask). The model accepts user form input and provides eligibility results immediately, with improved user experience and operational efficiency.



## 8. RESULTS AND DISCUSSION

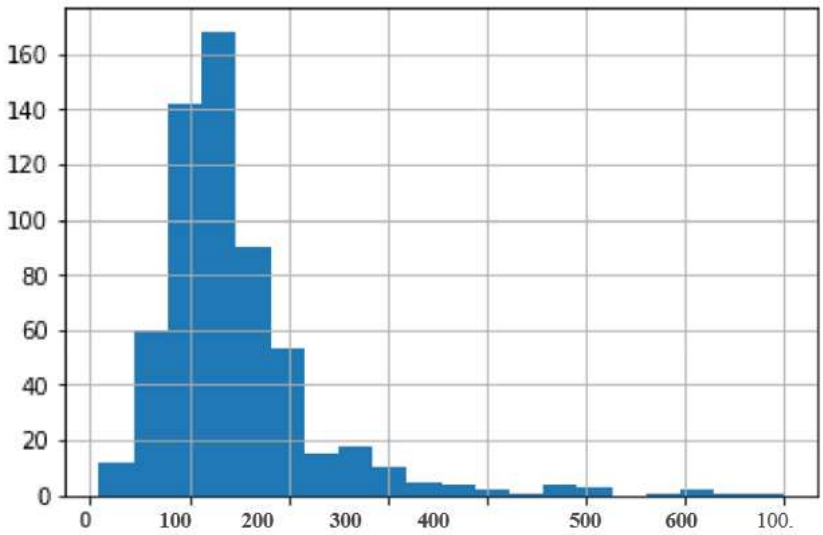



The above Box Plot also verifies the existence of many outliers/ extreme values. This is due to the income inequality in the society.




**Gender**

LoanAmount contains missing as well as extreme values, whereas ApplicantIncome contains a few ex treme values.




The extreme values are realistic, i.e. certain individuals may apply for high value loans because of certain requirements. So rather than treating them as outliers, let's attempt a log transforma tion to neutralize their influence:

**Conclusion**

This research effectively utilized machine learning methods to automate the loan eligibility prediction task. Through the use of a dataset with applicant demographic and financial data, we created a logistic regression model that could effectively predict loan approval results. Preprocessing operations such as missing value handling, categorical variable encoding, and new feature engineering like TotalIncome were instrumental in improving model performance.

The ultimate model had a precision of around 80.95%, validating its ability to detect patterns determining loan approval judgments. Of all the features, Credit_History, Education, and TotalIncome were identified to be strong determinants of qualification. These observations are consistent with actual expectations and offer useful suggestions for financial organizations looking to enhance their lending techniques.

Machine learning-based automation of loan eligibility checks not only enhances business efficiency but also reduces human bias and accelerates decision-making. Future research could include incorporating ensemble learning techniques or deep learning algorithms to further improve accuracy and responsiveness. Further, implementing the model in a real-time application system could offer an uninterrupted experience for both financial institutions and applicants.

In summary, machine learning presents a robust and scalable solution for making traditional loan processing an intelligent, data-driven system.

**REFERENCES**


[1] Al Mamun, M., Farjana, A., &Mamun, M. (2022): Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis. 7th North American International Conference on Industrial Engineering and Operations Management, Orlando, Florida, USA, June 12-14.

[2] Awodele, O., Alimi, S., Ogunyolu, O., Solanke, O., Iyawe, S., & Adegbie, F. (2022, November). Cascade of Deep Neural Network And Support Vector Machine for Credit Risk Prediction. In 2022 5th Information Technology for Education and Development (ITED) (pp. 1-8). IEEE.

[3] Bansal, M., Goyal, A., &Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal, 3, 100071.

[4] Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020, December). Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.

[5] Kadam, A. S., Nikam, S. R., Aher, A. A., Shelke, G. V., &Chandgude, A. S. (2021). Prediction for loan approval using machine learning algorithm. International Research Journal of Engineering and Technology (IRJET), 8(04).

[6] King, T., &Frishberg, I. (2001). Big Loans, Bigger Problems: A Report on the Sticker Shock of Student Loans. Kumar, C. N., Keerthana, D., Kavitha, M., &Kalyani, M. (2022, June).

[7] Customer Loan Eligibility Prediction Using Machine Learning Algorithms in Banking Sector. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 1007-1012). IEEE.

[8] Mohankumar, M., Amuthakkani, S., &Jeyamala, G. (2016). Comparative analysis of decision tree algorithms for the prediction of eligibility of a man for availing bank loan. Age, 19, 60.

[9] Nayak, D. S. K., Routray, S. P., Sahooo, S., Sahoo, S. K., & Swarnkar, T. (2022, August). A Comparative Study with Next Generation Sequencing Data and Machine Learning Approach for Crohn's Disease (CD) Identification. In 2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS) (pp. 17-21). IEEE.