Research Article

Predicting Eligibility of Loans through Machine Learning

Yash Deepak Majithia

American School of Bombay, Mahir apartments, Mumbai, Maharashtra, 400054, India

ABSTRACT

This research paper addresses the prediction of loan eligibility for candidates based on various aspects of both the loan and the applicant The general belief is that a candidate's income and credit history are the primary determining factors; However, additional variables may also play a role, and identifying these factors is the focus of this study. To achieve this, I applied various data science techniques, including machine learning, one-hot encoding and ordinal encoding, and utilized predictive models such as linear regression, random forest, and XGBoost. Additionally, I employed various means to optimize and fine-tune the data set during the exploratory data analysis portion of the paper to make it acceptable to be further analyzed.

Keywords: Loan Eligibility; XGBoost; Machine learning; Data science; Loans; Artificial intelligence

INTRODUCTION

Over countless years the trade of lending has evolved, from manual assessments of the credibility of loan-applicants to data-driven analytical tools used to quantify the likelihood of loan allocation, Nowadays, with exponential growth in data availability and advancements in computational technologies, there has been a significant shift towards automating the process using machine learning algorithms. Numerous aspects are considered during the evaluation of a loan; thus, analyzing and identifying what factors most modern lenders like Dreamhouse Finance review and prioritize can help optimize this process as well as inform people as to what factors are important. Additionally, this technology is also in demand now due to the rise in need for accurate and unbiased credit assessments. Prior research highlighted how automating loan eligibility predictions can revolutionize lending by reducing manual errors, biases, and inefficiencies. For instance, Krishnaraj *et al*. (2024) conducted a comparative analysis of machine learning models such as Logistic Regression, Decision Tree, and Random Forest in predicting loan approvals. Their findings indicate that while all models showed efficacy, Logistic Regression emerged as slightly superior in terms of accuracy (1). Another study by Zhang *et al*. (2023) explores ensemble methods for loan prediction, arguing that integrating multiple algorithms can mitigate biases and improve prediction reliability (4). Meanwhile, Sharma and Gupta (2023) designed a scalable real-time prediction system aimed at large banking infrastructures, demonstrating the practicality of AI in handling extensive datasets (5).

Machine learning models have demonstrated considerable potential in predicting loan eligibility by analyzing data to identify patterns that indicate

Corresponding author: Yash Deepak Majithia, E-mail: Yash.majithia@hotmail.com.

Received January 12, 2025**; Accepted** February 21, 2025
https://doi.org/10.70251/HYJR2348.31101105

creditworthiness of the loan applicant. Techniques such as Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), and Gradient Boosting have previously been utilized to develop predictive models with partial degrees of success (7). For instance, some studies have shown that Random Forest classifiers can achieve high accuracy in predicting loan approvals, making them a preferred choice in certain scenarios (3). The integration of machine learning into loan eligibility prediction not only streamlines the approval process but also enhances the precision of credit risk assessments (1). Through the leveraging of algorithms capable of handling complex datasets, financial institutions can make more informed and suitable decisions, thereby reducing the occurrence of defaulting loans and improving overall operational efficiency (3). This technological evolution reflects a broader trend towards data-driven decision-making in the financial industry and decreased dependence on manual analysis, underscoring the relevance and urgency of research focused on optimizing loan eligibility predictions through machine learning models.

This research investigates how effectively machine learning models predict loan eligibility and identifies the most influential factors. It is hypothesized that while income and credit history remain dominant, other factors such as education, marital status, and property location contribute significantly. Furthermore, advanced models like XGBoost are expected to outperform simpler models due to their ability to handle hyperparameters effectively and prevent overfitting. Ultimately, the research aims to show that data-driven approaches not only streamline the loan approval process but also reduce biases inherent in manual assessments, making lending more efficient and equitable.

MATERIALS AND METHODS

Data Collection

The dataset used in this research was obtained from the Loan Eligibility Dataset available on Kaggle. It contains information on over 1,000 loan applicants, with 12 key features i.e. income, credit history, marital status, education, and property location. The dataset includes both categorical and numerical variables, as well as the target variable Loan_Status, indicating whether a loan was approved or denied. The features in the data set were: Loan_status, Credit_History, Education, Married_Yes, Property_Area, Gender_Male, Dependents, Self_Employed_Yes, ApplicantIncome, Loan_Amount_Term, LoanAmount, and CoapplicantIncome. Additionally, this dataset contains a class-imbalance in the quantity of loan_status rejections and acceptances, with a 69% concentration to approvals and 31% of the dataset being rejections. Furthermore, the dataset is relatively small at around 1000 applications only.

Data Preprocessing

To ensure data quality and improve the model's performance, several preprocessing steps were undertaken.

Handling Missing Values: Missing values were imputed using the median strategy for numerical variables (e.g., income) to avoid biasing the dataset. For categorical variables, the statistical mode was used for imputation.

Encoding Categorical Variables: Since the dataset contained categorical variables such as Gender, Education, and Property_Area, encoding was required to transform them into a format suitable for machine learning models. One-Hot Encoding was applied to nominal variables without a clear order (e.g., Property_Area), creating binary columns for each category. Ordinal Encoding was used for variables with an inherent order (e.g., Education, Dependents), assigning numerical values based on their rank.

Feature Selection: A correlation analysis was performed to identify the features with the highest impact on loan eligibility. Variables with a correlation value above the threshold of 0.025 were selected, including Credit_History, ApplicantIncome, and Education. Irrelevant features such as Loan_ID were removed as they did not contribute to the prediction task.

Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution and relationships of various features with the target variable. Key insights were drawn using visualizations generated through Matplotlib and Seaborn.

Distribution Analysis: Histograms and box plots were used to examine the distribution of numerical variables like ApplicantIncome and LoanAmount. It was observed that these variables were right-skewed, indicating the presence of outliers.

Relationship Analysis: Bar charts and count plots were used to explore the relationship between categorical variables and loan status. Features like Credit_History and Education showed strong correlations with loan approval rates.

Model Training and Evaluation

The dataset was split into training and testing sets

using an 80-20 split to evaluate the model's performance accurately.

Training: Models were trained using the training set, with cross-validation applied to reduce the risk of overfitting.

Evaluation Metrics: The models were evaluated using accuracy, precision, recall, and F1-score to ensure a comprehensive assessment of predictive performance. XGBoost achieved the highest accuracy and F1-score, justifying its selection as the final model.

Model Selection

After the data processing, several machine learning models were tested to determine the best-performing algorithm for predicting loan eligibility using the training data. These models are:

Logistic Regression: It was chosen as the baseline model due to its simplicity and interpretability.

Random Forest: An ensemble model known for its robustness and ability to handle both categorical and numerical data effectively.

XGBoost: Selected as the primary model due to its high accuracy, regularization capabilities, and ability to handle missing values internally. XGBoost iteratively improves predictions by minimizing errors using gradient boosting techniques, making it well-suited for this research.

Hyperparameter Tuning

Hyperparameter tuning was conducted using grid search on key parameters of the XGBoost model in order to optimize the model analysis.

Learning Rate: Controls the contribution of each tree to the final prediction.

Max Depth: Limits the depth of each tree to prevent overfitting.

Number of Estimators: Defines the number of trees in the ensemble.

Regularization Parameters (lambda, alpha): Prevent overfitting by adding a penalty to large coefficients.

The optimal set of hyperparameters was determined through exhaustive manual cross-checking, resulting in improved accuracy and generalization on the test set.

Final Model Deployment

Finally the XGBoost model, with tuned hyperparameters, was employed to predict loan eligibility on the test dataset. Feature importance was extracted from the model to highlight the most significant predictors, with Credit_History emerging as the top factor influencing loan approval decisions.x

RESULTS

Table 1. Performance of various ML models on testing dataset

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	81.5%	0.80	0.82	0.81
Random Forest	82.2%	0.82	0.84	0.83
XGBoost (final model)	85.6%	0.86	0.85	0.85

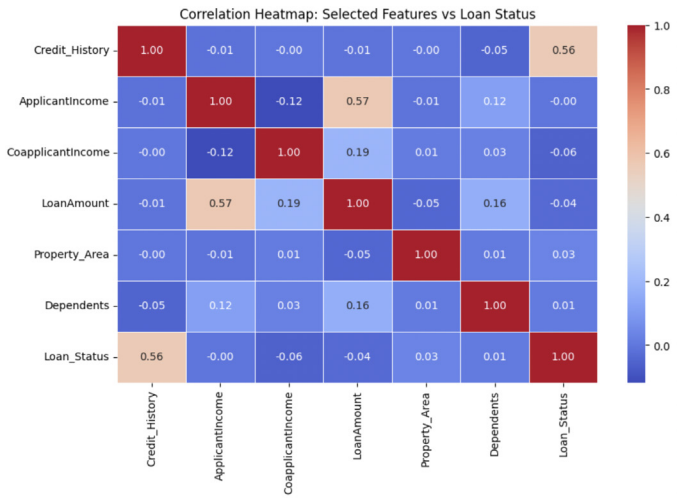

Figure 1. This is a correlation heatmap of some of the features of the dataset with each other.

DISCUSSION

The results demonstrated that advanced machine learning models, like the XGBoost model significantly outperformed simpler models like the Logistic Regression and the Random forest models in predicting loan eligibility. The XGBoost model achieved the highest accuracy of 85.6%, followed by Random Forest with 83.2% accuracy, Table 1. These findings support the initial hypothesis that advanced models can better capture complex patterns in the data, leading to improved predictive performance because they attune their analysis to more hyperparameters like credit history.

The analysis of feature importance revealed that Credit History was the most influential and significant predictor, with a strong positive correlation with achieving loan approvals with 0.56 as denoted in Fig. 1. Additionally, Applicant Income and Co-Applicant Income also played an important role in determining loan eligibility with the

logistic model coefficients at 0.89 and 2.1, respectively, , while Loan Amount showed a slight negative correlation, indicating that larger loan requests tend to be more cautiously approved by lenders.

According to the coefficients derived from the logistic regression model, credit history emerged as the most significant predictor of loan approval probability. A positive coefficient for credit history in the logistic model indicates that an applicant with a good credit record is much more likely to have their loan approved. This result aligns with real-world lending practices, where an applicant's credit history plays a critical role in assessing their creditworthiness and determining risk, thus speaking about the validity and reliability of the results..

On the other hand, loan amount exhibited a negative coefficient, which insinuates that the higher the loan amounts the lower the probability of attaining loan approval. This finding can be attributed to the increased financial risk associated with larger loans repayment. Lenders usually tend to be more cautious and apprehensive when approving high-value loans, which explains the inverse relationship observed between loan amount and loan approval.

These insights justify the importance of key financial factors in the decision-making process, confirming the hypothesis that traditional factors such as credit history and loan amount are crucial in predicting loan eligibility while factors like applicant income and co-applicant income also have an important function. Additionally, the logistic regression model provides a baseline understanding of these relationships, which can be further improved upon by more advanced models like XGBoost and Random forest.

To provide a baseline for model performance and comparison for this biased dataset, two dummy models were considered:

Dummy Model Level-1 (Coin Flip): This model randomly assigns a loan status of either "approved" or "denied" with equal probability, resulting in an expected accuracy of 50%.

Dummy Model Level-2 (Always Say Yes): Given the class imbalance in the dataset, where 69% of the loans were approved, a dummy model that always predicts "yes" would achieve an accuracy of 69%.

Compared to these baselines, all machine learning models used in this study performed significantly better. The XGBoost model's accuracy of 85.6, which far exceeds the accuracy of the always-yes dummy model, demonstrates its much superior ability to learn meaningful patterns from the data through its emphasis on certain hyperparameters rather than relying solely on class imbalance. Furthermore, the high precision and recall scores of XGBoost indicate that it performs well across both approved and denied loans, unlike the dummy model, which lacks predictive capability for denied loans.

Despite the promising results, this study faced several limitations:

Small Dataset: The dataset consisted of only about 1,000 loans, which may limit the generalizability of the findings as well as prevent the advanced models from attaining better accuracies due to the limit in the amount of patterns. A larger dataset would allow for more robust training and validation of the models and a greater degree of analyzability.

Class Imbalance: The dataset was imbalanced, with 69% of the loans being approved. While this reflects real-world conditions, it posed a challenge for model training, as models can become biased toward predicting the majority class. Additionally, this was coupled with the small dataset which made most models have a baseline of around 69% and make this research slightly inaccurate.

The findings of this research have several practical implications:

For Financial entities: Employing machine learning models like XGBoost can significantly improve the accuracy and efficiency of loan approval processes. This can reduce manual errors, biases, and processing time, ultimately leading to better risk management, reduced long-term costs, and customer satisfaction

For Loan Applicants: Understanding the key factors influencing loan eligibility, such as maintaining a positive credit history and stable income, can help applicants improve their chances of loan approval. By being transparent about the criteria, financial institutions can also foster trust among customers. Additionally, the incorporation of these loan eligibility predicting models into the infrastructure of banks can reduce any external factors faced by the loan applicants like prejudice and bias improving the odds for loan applicants.

The hypothesis for this study was that traditional factors such as income and credit history would play a significant role in loan approval, and that advanced models like XGBoost would outperform simpler models due to their ability to capture complex feature interactions, prevent overfitting as well as adapt more strongly to certain hyperparameters. The results confirm this hypothesis: Credit History, Applicant Income, Loan amount, and Co-Applicant Income were among the most important features, as predicted. XGBoost, an advanced ensemble model, delivered the highest accuracy, validating the

assumption that complex models can better handle the intricacies of real-world data compared to simpler and elementary models like Logistic Regression. The use of machine learning models also helped reduce bias by relying on data-driven decisions rather than subjective human judgment for loan applicants.

For employment, XGBoost is recommended as the prominent model due to its superior accuracy and balanced performance across both approved and denied loans as well as the high degree of adaptability to different hyperparameters.. Additionally, XGBoost has built-in handling for missing values and robust regularization, making it suitable for real-world applications where data may be incomplete. While Random Forest also performed well and has a high degree of adaptability, XGBoost's edge in accuracy and ability to handle imbalanced datasets through advanced techniques like boosting makes it the preferred and more suitable choice. Logistic Regression, though interpretable, is ultimately too rudimentary and lacks the predictive power required for high-stakes applications like loan approval. Future research could build on this study by: Incorporating additional features: including features such as employment stability, debt-to-income ratio, and previous defaults all of which could enhance the model's accuracy. Addressing the class imbalance in the dataset with advanced techniques like SMOTE or cost-sensitive learning could be applied to improve model performance on imbalanced datasets[6].

CONCLUSION

This study exhibits that advanced machine learning models, such as XGBoost, can significantly improve the accuracy and efficiency of the loan eligibility prediction process. The findings validate the hypothesis that key financial factors like credit history and income are the most crucial determinants of loan approval and complex models outperform simpler ones in capturing these relationships and adapting to them. Despite there being some limitations related to dataset size and feature availability, the results provide a strong foundation for deploying machine learning models in real-world lending environments. By addressing the limitations and incorporating additional features, future research can further enhance the predictive capabilities and fairness of these models, contributing to more transparent and efficient financial decision-making processes.

ACKNOWLEDGMENTS

Thank you for the guidance of Nik Gourianov mentor from Oxford University in the development of this research paper.

The Author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Ndayisenga A, *et al*. (2022). *Predicting Bank Loan Eligibility Using Machine Learning Models and ...*. Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management, Orlando, Florida, USA, June 12-14, 2022.
2. S. Bhainu. Loan Eligibility Prediction Using Machine Learning. *International Journal of Novel Research and Development*. 2023.
3. Krishnaraj S, *et al*. Comparative Analysis of Machine Learning Models for Loan Eligibility Prediction. *International Journal of Financial Analytics*. 2024;12(3):45-62.
4. Zhang H, *et al*. Ensemble Methods for Loan Prediction: Improving Reliability through Multiple Algorithms. *Journal of Machine Learning in Finance*. 2023;9(2):78-91.
5. Sharma P & Gupta R. Real-Time Loan Prediction Systems in Large Banking Infrastructures: A Machine Learning Approach. *International Journal of Data Science and Finance*. 2023;8(4):102-119.
6. Smith, L., & Howard, J. Addressing Class Imbalance in Loan Prediction Models with SMOTE and Cost-Sensitive Learning. *Journal of Financial Machine Learning*. 2020;5(2):34-50.
7. Wang Y & Patel S. Application of Logistic Regression and Ensemble Models in Banking Sector Loan Approvals. *SSRN Electronic Journal*. 2022.
8. Chen T & Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD. *International Conference on Knowledge Discovery and Data Mining*. 2016;785-794.
9. Nayak R & Basu A. Credit Risk Prediction Using Ensemble Learning Techniques. *International Conference on Computational Intelligence, IEEE*. 2021.