Predicting Loan Eligibility Approval Using Machine Learning Algorithms

Guangxuan Chen
School of Information Management and Engineering, Shanghai University of Finance and Economics,
Shanghai, 200433, China

Keywords: Loan, Finance, Machine Learning, Classification, AdaBoost.

Abstract: The survival and profitability of financial institutions are closely related to the recipients of the loan. However, traditional loan approval methods are struggling to keep up with the diversifying and rapidly growing loan applications. In this context, machine learning techniques present a promising solution. Previous studies have only attempted limited models, while this study aims to achieve higher loan approval prediction accuracy by comprehensively comparing multiple mainstream models. This article selects a publicly available dataset from Kaggle, conducts detailed data preprocessing, and comprehensively trains eight mainstream models. The evaluation metrics used are precision, accuracy, and F1-score. Among these models, AdaBoost performed the best, achieving the highest Accuracy (84.95%) and the best F1-score (0.8957). XGBoost performed the best in terms of Precision. This study presents a more accurate method for loan approval and demonstrates the reliability of machine learning in supporting intelligent financial decision-making. This tool helps financial institutions to efficiently and fairly assess the eligibility of loan applicants, streamline the loan approval process, and promote financial inclusion.

1 INTRODUCTION

Financial institutions act as intermediaries, channeling idle funds to individuals, enterprises, or projects needing funds. This improves the efficiency of resource utilization and promotes progress in all walks of life. Loans, as a core business of financial institutions, play a crucial role in promoting economic growth. To improve the loan approval process, it is necessary to develop new methods that are more efficient and effective. Traditional methods for approving loans rely too heavily on credit scoring models, which are often overwhelmed by the need to respond to diverse loan applications, consider emerging financial businesses, and handle large amounts of data. In the meanwhile, only professionals can assess the qualifications and loan default risks associated with applicants using these methods. In today's complex and competitive financial markets, traditional methods often come with inefficiencies and uncertainty in approval decisions. Gupta et al. state that traditional loan approval is a difficult and risky process (Gupta et al. 2020).

Loan approval is the process of distinguishing creditworthy applicants from potential defaulters. It is important to ensure both efficiency and fairness. However, traditional loan approval methods may be hindered by factors such as human subjective judgment, which can exclude some potential borrowers from eligibility. As the concept of financial inclusion grows, it is imperative to ensure that every potential borrower receives fair and speedy loan approvals; relative study shows that financial inclusion can help reduce poverty, increase financial innovation, strengthen the stability of the financial sector, and improve the economy (Ozili 2021).

Fortunately, most financial institutions are digitally transformed today. The integration of digital technology in the financial industry has led to new forms of digital finance. Machine learning (ML) technology has opened up new possibilities for loan approval. By using large-scale datasets and advanced complex algorithms for model training, ML can speed up the approval process and help financial institutions accurately assess the credit risk of loan applicants. It also improves the accuracy and fairness of approvals, achieving more universal financial inclusion.

Its effects and limitations in loan qualification approvals require further research. Previous research suggests that banks should use a combination of

Chen, G.
Predicting Loan Eligibility Approval Using Machine Learning Algorithms.
DOI: 10.5220/0012828200004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 512-517
ISBN: 978-989-758-690-3

multifaceted customer attributes when deciding to grant a loan using ML models (Sheikh et al. 2020). Researchers have investigated the application of several basic models, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest (RF), and have achieved high accuracy in predicting loan approvals (Singh et al. 2021, Tumuluru et al. 2022, Saini 2023).

However, most existing research has only tested a limited number of models. This paper employs a meticulous data preprocessing technique and conducts a thorough comparison of various popular models to achieve high accuracy.

By introducing ML models, this paper offers financial institutions a fast, straightforward, and unbiased method to screen qualified loan applicants. They can make data-driven decisions, improve the efficiency and accuracy of loan placement, and reduce manual intervention through the automation of loan eligibility approval. This is an opportunity for financial institutions to move towards a smarter future while also promoting financial inclusion.

This paper adopts the same original dataset on Kaggle as several references refines the data preprocessing work through feature analysis, and comprehensively trains various mainstream models. The aim is to compare the strengths and weaknesses of existing methods more intuitively, which will help in selecting the optimal solution.

2 METHOD

2.1 Data Collection

A dataset was collected from Kaggle, containing 11 applicant characteristics and their final application status (Loan Prediction Problem Dataset 2019). Prior research has been conducted using this dataset. For example, Uddin et al. constructed an ensemble learning model by using the voting method to combine the three best-performing ML models out of the nine they tested, achieving an accuracy rate of up to 87.26% (Uddin et al. 2023). In contrast, Orji et al.'s experimental results showed that Random Forest outperformed the other five models they used in terms of accuracy score (Orji et al. 2022). Additionally, the results of Mridha et al. showed that logistic regression had the highest accuracy of 80.43% (Mridha et al. 2022). The previous research on this dataset indicates that it is valuable and representative for studying the prediction of loan approval.

2.2 Exploratory Data Analysis

2.2.1 Data Overview

The original dataset comprises thirteen fields, some of which have missing values. A summary of the original dataset is shown in Table 1.

The 'Loan ID' field serves as the primary key, which means that each record is unique. This implies that the dataset contains no duplicates, but the primary key is not useful for further research and should be considered for deletion.

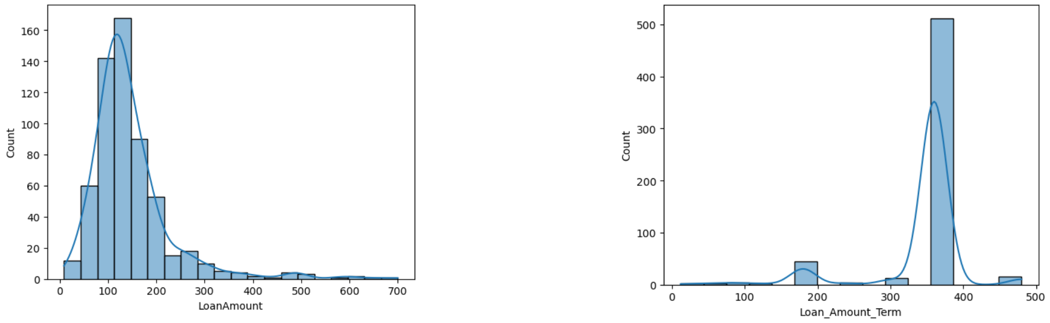
Table 1**:** Overview of the dataset.

Attribute Name	Non-Null Count	Type
Loan ID	614	String
Gender	601	String
Married	611	String
Dependents	599	String
Education	614	String
Self-Employed	582	String
Applicant Income	614	Integer
Co-applicant Income	614	Integer
Loan Amount	592	Integer
Loan Amount Term	600	Integer
Credit History	564	Integer
Property Area	614	String
Loan Status	614	String

Furthermore, there are missing values in the fields 'Gender', 'Married', 'Dependents', 'Self-Employed', 'Credit History', 'Loan Amount', and 'Loan Amount Term', and different treatment strategies are employed based on the characteristics of each field.

2.2.2 Univariate Analysis

Histograms were plotted for numerical variables with missing values to observe their distribution and characteristics (Fig. 1).



(a) Loan Amount

(b) Loan Amount Term

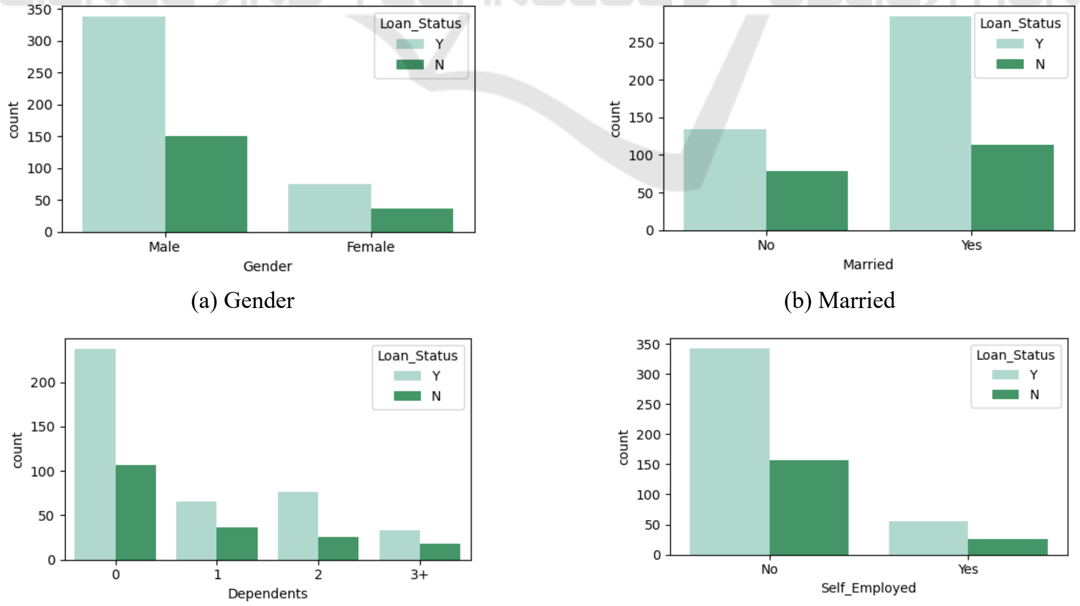
Figure 1: Histograms of numerical variables with missing values (Photo/Picture credit: Original).

It was found that the values of 'Loan Amount Term' were concentrated at 360 (>83%), indicating that this variable does not differentiate between different applicants and is not useful for subsequent prediction. Therefore, it was decided to delete it directly.

'Loan Amount' is left-skewed and not normally distributed, so it is appropriate to use the median to fill in the missing values.

2.2.3 Bivariate Analysis

This section focuses on examining the statistical relationship between the categorical variables with missing values and loan status. Several bar charts are plotted in Figures 2 & 3.



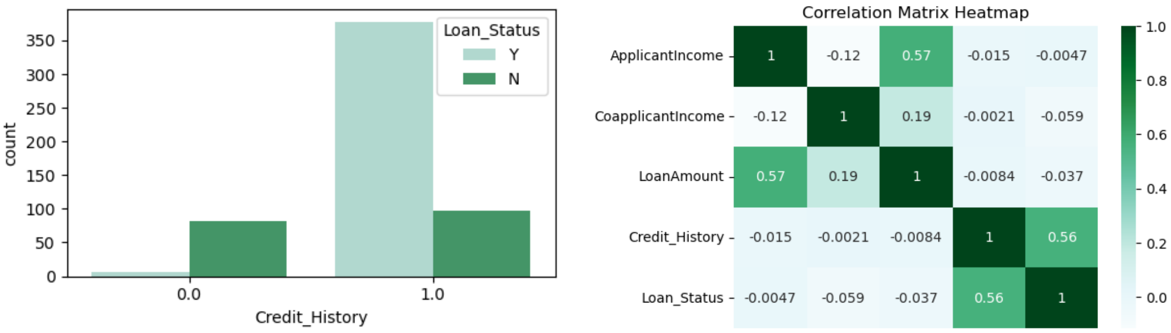
(a) Gender

(b) Married

(c) Dependents

(d) Self-employed

Figure 2: The distribution of loan status across categorical variables (Photo/Picture credit: Original).



(a) Credit History

(b) Correlation Matrix Heatmap

Figure 3: Distribution of loan status by credit history; correlation matrix heatmap (Photo/Picture credit: Original).

Figure 3 demonstrates a strong correlation between 'Credit History' and 'Loan Status', as different values of the former lead to opposite proportions of the latter. The calculation of the correlation matrix also reflects this, so it would be a good idea to delete the record with the missing 'Credit History' value instead of blindly populating it. Blind filling can largely destroy the correlation, resulting in poorly fitted prediction models.

In the meantime, the proportions of 'Loan Status' are similar across the values of the other variables that have missing values. Therefore, the mode was chosen to fill them.

2.2.4 Discrete Attribute Encoding and Dataset Splitting

The categorical variables of the dataset have been numerically encoded for model training. The 'Gender' column has been transformed by assigning the value 0 to 'Female' and 1 to 'Male'. Similarly, the 'Married' and 'Self-Employed' columns have been converted to numeric representations, with 'No' corresponding to 0 and 'Yes' corresponding to 1. Additionally, the 'Education' column has also been transformed, where 'Not Graduate' is now represented by 0 and 'Graduate' by 1. In the case of the 'Dependents' column, the string category '3+' has been replaced with the numerical value 3 to simplify analysis. The 'Property Area' column, which was originally categorical, has been replaced with dummy variables.

2.3 Modeling

After these preprocessing steps, the dataset was thoroughly cleaned. Finally, the dataset has been divided into two separate parts - an 80% training set and a 20% test set.

This study explores a binary classification problem using eight well-known ML models: SVM, KNN, LR, Decision Tree (DT), RF, XGBoost, AdaBoost, and Gradient Boosting (GBDT). The effectiveness of these classifiers is thoroughly evaluated across multiple dimensions.

In the context of financial institutions, significant losses can occur when loans are extended to potential defaulters. Conversely, profitability depends on accurately identifying and approving loans for creditworthy applicants. In light of this, the study emphasizes the importance of positive predictive outcomes. The evaluation metrics used in this study are Precision, Accuracy, and F1-score. These metrics were chosen to comprehensively assess the performance of the classifiers in predicting loan approvals.

3 RESULTS

Table 2 summarizes the Precision, Accuracy, and F1-Score of each model.

Table 2: Precision, Accuracy, and F1-Score.

Model Name	Precision	Accuracy	F1-Score
SVM	0.681416	0.681416	0.810526
KNN	0.679612	0.646018	0.777778
LR	0.837209	0.831858	0.883436
DT	0.837209	0.831858	0.883436
RF	0.839080	0.840708	0.890244
XGBoost	**0.871795**	0.831858	0.877419
AdaBoost	0.848837	**0.849558**	**0.895706**
GBDT	0.837209	0.831858	0.883436

Among them, the highest Precision on the test set is 87.18% for XGBoost, followed by 84.88% for AdaBoost. AdaBoost also achieved the highest accuracy (84.95%) and F1 score (0.8957).

Considering all evaluation metrics, the best-performing model is AdaBoost.

4 DISCUSSION

Essentially, the prediction problem is a supervised learning task that uses ML models to implicitly capture the underlying patterns in the manual approval process.

The optimal model obtained in this study is AdaBoost, which is consistent with the choice made by previous researchers (Kumar et al. 2022). Specifically, the performance of this model is significantly better than the results of Mridha et al (Mridha et al. 2022). This is due to the ensemble learning algorithm's excellent generalization capabilities, which were not examined by Mridha et al. Additionally, the data preprocessing process differs between the two studies.

The author notes that XGBoost achieved 99% accuracy on the training set. This indicates that the model suffered from overfitting during training, which explains why the DT-based model while performing best on Precision, does not perform well on the metrics of Accuracy and F1-score. One possible explanation is that the training set was too small.

Compared to previous studies, this paper's advantage lies in its finer and more reasonable data preprocessing. This can be seen from the fact that the author's model still outperforms Mridha et al.'s model in terms of logistic regression (Mridha et al. 2022).

Due to the shortcomings of this paper, and the small size of the dataset, some models face the challenge of balancing overfitting and underfitting during parameter tuning. Future studies should select a more appropriate dataset to fully explore the intrinsic relationships among these variables.

5 CONCLUSION

This paper applies ML techniques to the prediction of loan approval outcomes. When predicting loan approvals, financial institutions typically focus on positive examples. Whether these borrowers repay on time and with interest determines whether the financial institution will make a profit or a loss. Appropriate preprocessing of the collected dataset was performed and eight models were trained, including SVM, KNN, LR, DT, RF, XGBoost, AdaBoost, and GBDT. The performance of the models was evaluated using Precision, Accuracy, and F1-score. The author concluded that the DT-based learning method AdaBoost produced the best prediction results, with a remarkable accuracy of 84.95%.

This study demonstrates the potential of ML in the financial sector, specifically in reducing costs and increasing efficiency. An example of this is the possibility of building an automatic loan approval system using the AdaBoost model in the future. This tool can help financial institutions review loan applications efficiently and fairly, streamline the loan approval process, and promote financial inclusion while improving efficiency.

Also, this research presents an innovative method for more accurate loan approval in the financial sector and provides reliable support for intelligent financial decision-making. The finding extends the empirical research in financial technology and provides valuable insights for optimizing decisions during the current digital transformation in the financial sector.

REFERENCES

A. Gupta, V. Pant, S. Kumar, P. K. Bansal, “Bank Loan Prediction System using Machine Learning,” in *2020 9th International Conference System Modeling and Advancement in Research Trends*, (IEEE, 2020), pp. 423–426.

P. K. Ozili, Forum for Social Economics **50** (4), 457-479 (2021).

M. A. Sheikh, A. K. Goel, T. Kumar, “An approach for prediction of loan approval using machine learning algorithm,” in *2020 International Conference on Electronics and Sustainable Communication Systems*, (IEEE, 2020), pp. 490–494.

V. Singh, A. Yadav, R. Awasthi, G. N. Partheeban, “Prediction of modernized loan approval system based on machine learning approach,” in *2021 International Conference on Intelligent Technologies*, (IEEE, 2021), pp. 1–4.

P. Tumuluru, L. R. Burra, M. Loukya, S. Bhavana, H. CSaiBaba, N. Sunanda, “Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms,” in *2022 Second International Conference on Artificial Intelligence and Smart Energy*, (IEEE, 2022), pp. 349–353.

P. S. Saini, A. Bhatnagar, L. Rani, “Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms,” in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering*, (IEEE, 2023), pp. 1821–1826.

Loan Prediction Problem Dataset, 2019, available at https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset.

N. Uddin, M. K. Uddin Ahamed, M. A. Uddin, M. M. Islam, M. A. Talukder, S. Aryal, International Journal of Cognitive Computing in Engineering **4**, 327–339 (2023).

U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu, P. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility," in *Proceedings of the 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development*, (IEEE, 2022), pp. 1–5.

K. Mridha, D. Barua, M. M. Shorna, H. N. Nouman, M. H. Kabir, A. V. Singh, "Credit Approval Decision using Machine Learning Algorithms," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, (IEEE, 2022), pp. 1–6.

C. N. Kumar, D. Keerthana, M. Kavitha, M. Kalyani, "Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector," in *7th International Conference on Communication and Electronics Systems*, (IEEE, 2022), pp. 1007–1012.