

The Comprehensive Analysis of Bank Loan Approval Prediction Based on Machine Learning Models

Deyi Li ^a

Boston University, One Silber Way, Boston, MA 02215, U.S.A.

Keywords: KNN, SVM, Decision Tree, Random Forest, XGboost.

Abstract: In contemporary society, given the current economic depression, many households are experiencing significant financial pressure, making it increasingly challenging to meet large capital requirements. The price and cost of products is increasing, and the capital required for a lump-sum purchase has become prohibitively high. Therefore, obtaining a bank loan represents one of the most practical solutions to address this need. However, for banks, the volume of daily loan applications is substantial, making it impractical to approve every request. This necessitates the allocation of limited funds to reliable applicants who have a strong likelihood of repayment. Most banks consider a customer's credit score as a critical factor in loan approval. However, when credit scores are unavailable in customer information, the alternative relevant data can be utilized to assess risk. Five models of machine learning, in this study, were applied to predict loan issuance status in both scenarios: including and excluding credit scores. The accuracy of prediction, precision of prediction, recall score of prediction, f1 and Area Under Curve (AUC) score of these models were compared and evaluated. When credit scores are available, the decision tree model attains the highest accuracy and AUC score compared to other models; among predictions made without credit scores, the random forest model provides the best comprehensive performance.


1 INTRODUCTION

In recent years, the loan industry has experienced rapid growth alongside the swift economic expansion. The development of the loan business has a stimulating effect on improving domestic demand and promoting consumption. The amount of personal loans continues to grow. According to the statistical data released by the Central Bank, as of the end of 2022, the balance of RMB loans extended by financial institutions stood at 213.99 trillion yuan, representing an 11.1% year-on-year increase. For the entire year, RMB loans expanded by 21.31 trillion yuan, marking a year-on-year rise of 1.36 trillion yuan. The risk problems such as credit default are also increasing, which partly limits the healthy development of the credit market (Tumuluru et al., 2022). Non-performing loans pose significant risks to banks' financial health and operational stability. An excessive accumulation of such loans can severely impair a bank's ability to function effectively. (Sheikh et al., 2020) Moreover, from a societal perspective,

non-performing loans can have detrimental effects, leading to a cascade of adverse consequences.

Due to the increasing number of bank loan applications, banks need to be very strict in the comprehensive examination of all aspects of the applicant. During this review process, banks must accurately assess whether applicants meet the eligibility criteria for loan approval based on the available information, thereby mitigating potential risks. (Anand et al., 2024) To realize an automatic process, reduce labor costs, and minimize risk exposure, banks can leverage machine learning models to build predictive systems (Khan et al., 2021). As a subset of the technology that can perform tasks typically requiring human intelligence, machine learning, is capable of rapidly processing and analyzing vast datasets, enabling it to make informed predictions based on historical data. (Uddin et al., 2023).

A considerable body of research has explored this topic as well. For instance, in research, the author employed a range of machine learning algorithms,

^a <https://orcid.org/0009-0009-1172-3435>

such as Random Forest, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression, to forecast the rate of loan approval eligibility and the random forest algorithm yielded the most favorable outcomes, achieving 81% accuracy (Tumuluru et al., 2022). Similarly, both paper and research evaluated the performances of these models and reported random forest is better than another model, achieving 80% - 90% accuracy (Khan et al., 2021; Sarkar et al., 2024).

The main objective of this research is utilizing an extensive dataset encompassing customer data, including credit score, annual income, loan amount, loan term, various kinds of assets evaluations, etc. to compare and evaluate the performance of various kinds of models for predicting loan issuance under two distinct feature combinations, to select the optimal model that enhances audit efficiency and accuracy while mitigating associated risks. These predictive models not only help applicants reduce waiting times for applications, but also assist banks in minimizing risks and accurately identify customer categories.

2 METHODOLOGY

2.1 Dataset Description

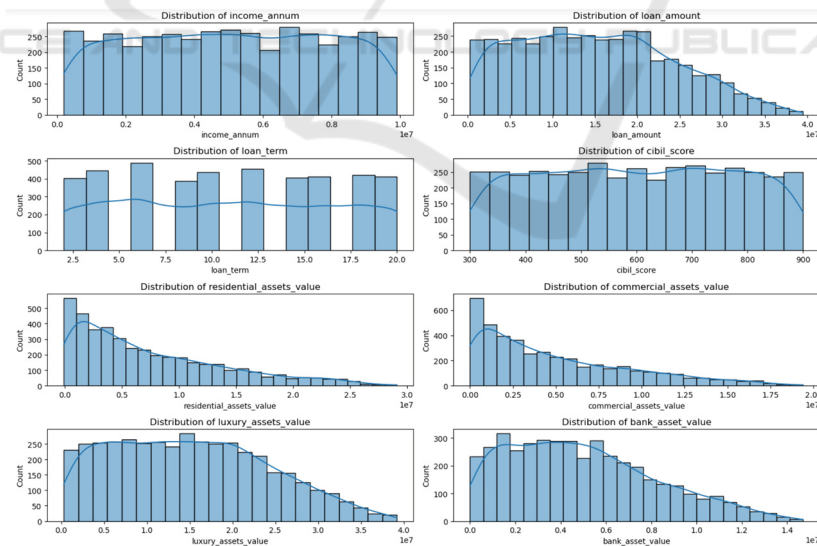


Figure 1: Numeric features distribution (Photo/Picture credit: Original).

This dataset named loan approval is from Kaggle. There are 4269 rows and 13 columns, including 10 numerical columns and 3 categorical columns, which are some detailed information of applicant customers, such as the number of their offspring, their income of every year, their amount of loan, their total term of loan, their credit score, the value of their different kinds of properties. The target column is the status of their loan, specifying whether each loan application was approved or rejected. The original dataset was split into two sets, a training set and a testing set, with 80% allocated to the training set. In this research, a training set was used to train the five different models based on all features mentioned above, and the rest of the features except for credit score to predict whether the customer in the testing set is approved for a loan or not.

2.2 Exploratory Data Analysis

Some of the numerical features have imbalanced distribution, such as the loan amount of those clients, the value of those assets, including residential, commercial, luxury, and bank property. Some of numerical features are balanced distribution, such as annual income, loan term, and credit score. Categorical features are almost balanced. As shown in two Figures 1, 2:

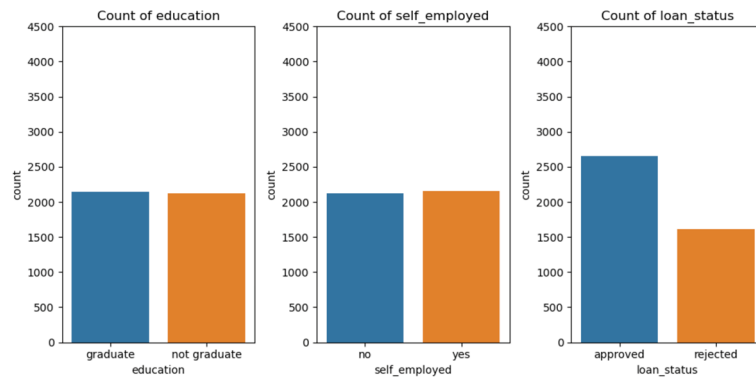


Figure 2: Categorical features distribution (Photo/Picture credit: Original).

2.3 Data Preprocessing

The paper converted categorical variables into numeric form. For example, in the education column, 1 represents graduate and 0 represents not graduate; in the column which indicates the status of employment, 1 represents that this person was self-employed and 0 represents that this person was not self-employed; in the column which indicates the status of loan, 1 represents approved and 0 represents rejected. The author also used standard scaler to scale some numerical features, such as the residential, commercial, luxury, and bank asset, these kinds of assets value of customers.

2.4 Machine Learning Models

2.4.1 K Nearest Neighbors (KNN)

KNN, a supervised learning algorithm, suitable for classification tasks, as well as regression tasks. (View of Prediction of Loan Approval in Banks Using Machine Learning Approach, 2025) The core concept of KNN is to identify the K samples in the training dataset that are most similar to a new data point based on a predefined distance metric. For classification tasks, the algorithm predicts the category of the new data point by analyzing the categories of its nearest K neighbors. And for regression tasks, it estimates the value of a new data point by calculating the average of the values of its nearest K neighbors.

2.4.2 Decision Tree (DT)

DT model, a kind of supervised learning algorithm, is usually employed in classification programs, as well as regression programs. This kind of predictive model is constructed upon a flowchart framework, which means that it makes choices based on input data. The

tree learning algorithm and supervised learning technology can be applied to the prediction model with high accuracy.

2.4.3 Support Vector Machine (SVM)

SVM, a kind of supervised learning algorithm, is primarily utilized for classification tasks, as well as regression tasks. SVM is particularly well-suited for small sample datasets and high-dimensional data.

2.4.4 Random Forest (RF)

RF is one of the prominent supervised learning methods mainly applied to machine learning problems involving regression and classification.

The core of this approach is that mixing different types of learning models together can enhance the aggregated results. The RF method provides tree-based prediction results and predicts the actual real model based on the prediction results that received the most votes. (Arun et al., n.d.) An increase in the number of trees within a forest generally leads to higher accuracy and reduces the likelihood of overfitting. (Alaradi & Hilal, 2020)

2.4.5 XGBoost

XGBoost is an efficient supervised learning algorithm suitable for classification and regression tasks. It can efficiently solve problems in the real world and is widely used in various competitions, so it is one of the most mainstream models today.

It is especially effective when dealing with structured data. (Yu et al., 2024) Its main idea is to construct a robust model for prediction by integrating multiple weak learners, thereby progressively reducing the prediction error of the model. XGBoost introduces several critical enhancements over traditional gradient boosting machines (GBM), which

substantially improve both the computational efficiency and predictive performance of the model.

3 RESULTS

3.1 Result of Prediction 1

The accuracy score of prediction, the precision of prediction, the recall of prediction, f1 score of

prediction, and AUC score of prediction for scenario 1 are calculated and shown in Table 1. Decision Tree model performed best in terms of Accuracy and AUC score, which means that Decision Tree model is generally correct in prediction and has the best discrimination power to distinguish between classes. SVM model has the lowest Accuracy and AUC score, which means that in this particular scenario, SVM model may exhibit inferior performance compared to alternative models.

Table 1: performance of models in scenario 1

Models	Accuracy score	Precision score	Recall score	F1 score	AUC score
KNN	0.954333	0.969466	0.956685	0.963033	0.953575
Decision Tree	0.982436	0.984962	0.986817	0.985889	0.981025
SVM	0.943794	0.980119	0.928437	0.953578	0.948739
Random Forest	0.975410	0.982955	0.977401	0.980170	0.974769
XGBoost	0.977752	0.979401	0.984934	0.982160	0.975439

3.2 Result of Prediction 2

The accuracy score of prediction, the precision of prediction, the recall of prediction, f1 score of prediction, and the AUC score of prediction for scenario 2 are calculated and shown in Table 2.

Overall, Random Forest model performed best in terms of Accuracy and AUC score, which means that Random Forest model is generally correct in prediction and has the best discrimination power to distinguish between classes. Decision Tree model has the lowest Accuracy and SVM model has the lowest AUC score.

Table 2: performance of models in scenario 2

Models	Accuracy score	Precision score	Recall score	F1 score	AUC score
KNN	0.576112	0.639209	0.730697	0.681898	0.526339
Decision Tree	0.558548	0.644195	0.647834	0.646009	0.529799
SVM	0.621780	0.621780	1.000000	0.766787	0.500000
Random Forest	0.606557	0.639885	0.839925	0.726384	0.531417
XGBoost	0.572600	0.637417	0.725047	0.678414	0.523514

4 DISCUSSION

Compared to the findings of other studies, the random forest model demonstrates superior performance as a relatively optimal approach for predicting loan eligibility. But, the dataset utilized for training the model in this experiment is constrained, and future enhancements could involve automating the model to facilitate the automatic incorporation of new data into the training database. This would contribute to more robust training and improved predictive performance. This research should also use k-fold cross-validation in the experiment to ensure that the predictions are stable and reliable (Khan et al., 2021). Loan forecasting is a complex and challenging problem in practice. Macroeconomic factors such as unemployment and inflation can influence loan

defaults. Additionally, incomplete training data may introduce biases, leading to unfair or inaccurate predictions. These factors are inherently unpredictable and difficult to quantify, necessitating more comprehensive data for model training and ongoing model optimization to ensure accurate and robust predictions (Sarkar et al., 2024).

5 CONCLUSION

In this research, it conducted a comprehensive comparison and analysis of machine learning models utilized for predicting loan approvals. The predictive process begins with data preprocessing, including converting the categorical features into numerical form, scaling the numerical features, and splitting the original dataset into two parts: the training set and the

testing set, as well as utilizing K-Nearest Neighbors (KNN), Support Vector Machine(SVM), Decision Tree, Random Forest, and XGBoost, these five models to learn and analyze the content of the training model in the two scenarios respectively, and then apply them to the testing set to predict whether to issue loans for a customer. Finally, the accuracy score of prediction, the precision score of prediction, the recall score of prediction, f1 score of prediction, and AUC score of prediction for scenario 1 and 2 are calculated and the score of accuracy, as well as AUC is used as evaluation factors to select the optimal model.

In the case of credit scores, the Decision Tree model exhibits the highest accuracy and AUC score, making it the optimal choice. In the case of absence credit score, the Random Forest (RF) model outperforms other models in the comparison of accuracy score and AUC values.

In the future, these models still need to be trained on the large and noisy dataset to improve their stability and accuracy. They will continue training and testing repeatedly to perfect them to help banks minimize the risk.

REFERENCES

- Alaradi, M. and Hilal, S. 2020. Tree-Based Methods for Loan Approval.
- Anand, R., Singh, H., Sardana, K., Gupta, D. N., Sindhwani, N. and Mittal, M. 2024. Loan Approval Prediction Using Machine Learning. *Lecture Notes in Networks and Systems*, 357–366.
- Arun, K., Ishan, G. and Sanmeet, K. (n.d.) Loan Approval Prediction based on Machine Learning Approach. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 79–81.
- Khan, A., Bhadola, E., Kumar, A. and Singh, N. 2021. Loan Approval Prediction Model: A Comparative Analysis. *Advances and Applications in Mathematical Sciences*, 20(3), 427–435.
- Sarkar, T., Rakhra, M., Sharma, V. and Singh, A. 2024. An Empirical Comparison of Machine Learning Techniques for Bank Loan Approval Prediction.
- Sheikh, M. A., Goel, A. K. and Kumar, T. 2020. An Approach for Prediction of Loan Approval using Machine Learning Algorithm. *IEEE Xplore*, 1 July.
- Tumuluru, P., Burra, L. R., Loukya, M., Bhavana, S., SaiBaba, H. M. H. and Sunanda, N. 2022. Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms. *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*.
- Uddin, N., Ahamed, M. K. U., Uddin, M. A., Islam, M. M., Talukder, M. A. and Aryal, S. 2023. An Ensemble Machine Learning Based Bank Loan Approval Predictions System with a Smart Application.

International Journal of Cognitive Computing in Engineering, 4, 327–339.

View of Prediction of Loan Approval in Banks using Machine Learning Approach. 2025 vandanapublications.com. Available at: <https://ijemr.vandanapublications.com/index.php/j/article/view/1318/1163> (Accessed: 10 March 2025).

Yu, K., Xia, S., Zhang, Y. and Wang, S. 2024. Loan Approval Prediction Improved by XGBoost Model Based on Four-Vector Optimization Algorithm. *Applied and Computational Engineering*, 82(1), pp. 35–44.