Predicting Loan Approval Decisions: A Comparative Machine Learning Study

Wanxuan Sun

Major in Big Data Management and Application, Harbin Institute of Technology, Harbin, China
2023110934@stu.hit.edu.cn

Abstract. The experiment used the public Kaggle Loan Approval Prediction Dataset, which contains 4,269 applications. The paper treated loan status as a two-class target and kept the preprocessing workflow deliberately simple: removing the identifier field, checking missing values, encoding categorical variables, splitting the data into training and test samples, and scaling features for algorithms that are sensitive to variable magnitude. Five models were compared, namely Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machine. Accuracy, precision, recall, and F1-score were used to evaluate the hold-out test results. The two tree-based models produced the clearest performance, and the Decision Tree obtained the highest score in this run, with an accuracy of 0.9813 and an F1-score of 0.9849. Random Forest was only slightly lower. Its feature importance output showed that CIBIL score dominated the model decision process, while loan term and loan amount contributed at a much smaller level. The findings suggest that machine learning can support loan screening on structured data, but the result should be understood as a classroom-style model comparison rather than a ready-to-use banking system.

Keywords: Loan approval prediction; Credit scoring; Supervised learning; Decision tree; Feature importance

1 Introduction

Loan approval is not only a technical classification problem; it is also a risk-control decision. A loan that is accepted too easily may later become a default loss, while a good applicant who is rejected may reduce customer trust and business opportunities. In many lending scenarios, reviewers still rely on a mixture of credit rules, previous experience, and manual checking. This approach is understandable, but it becomes slow and inconsistent when the number of applications is large and when the information behind each application is multidimensional.

Structured lending data makes it possible to test whether supervised learning can assist this decision process. Variables such as credit score, annual income, loan amount, loan term, number of dependents, and asset value do not work independently. For example, the same loan amount may imply different risk levels for applicants

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_155

with different incomes or credit histories. A machine learning model can learn such patterns from historical records and then provide a predicted approval outcome for new applications. In this sense, the model is not intended to replace human review, but it can provide an additional reference for screening and comparison.

Previous credit-scoring research provides two useful starting points for this project. First, Logistic Regression is still a necessary benchmark because it is simple, transparent, and widely introduced in standard machine learning texts together with Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) [1]. Khandani, Kim, and Lo showed that machine-learning-based credit-risk models could improve out-of-sample prediction compared with traditional scoring methods [2]. Second, later studies suggest that non-linear or ensemble methods often perform better when credit data contain more complicated relationships. Dastile, Celik, and Potsane reviewed statistical and machine learning credit-scoring studies and reported the frequent advantage of ensemble learning [3]. Dumitrescu et al. improved Logistic Regression by adding decision-tree effects, which also points to the value of modelling non-linearity [4].

Interpretability is another important issue in financial applications. A highly accurate model is less useful if its decision logic cannot be explained to reviewers or applicants. For this reason, recent work has paid attention to explainable modelling. de Lange et al. combined LightGBM and SHAP in a bank credit-assessment setting and linked prediction performance with feature-level interpretation [5]. Liu, Zhang, and Matousek compared several algorithms on Chinese online lending data and found Random Forest to be the strongest among the tested models [6]. Guo, Zhou, and Qian further showed that mobile application information can improve credit-risk evaluation when added to traditional financial variables [7]. Based on these studies, this study compares several mainstream classification models on a public loan approval dataset to evaluate their effectiveness in loan approval prediction.

To address this question, a public loan approval dataset was used, and five classification models were developed and compared: Logistic Regression, Decision Tree, Random Forest, KNN, and SVM. Through this research, this study aims to evaluate the effectiveness of different machine learning approaches in loan approval prediction and to identify the key factors that influence lending decisions.

2 Methodology

2.1 Dataset and Preprocessing

The dataset was taken from the Kaggle Loan Approval Prediction Dataset [8]. It includes 4,269 application records and 13 original fields. The target field is loan status, which records whether each application was approved or rejected. The remaining fields describe applicant and loan information, including number of dependents, education, self-employed status, annual income, loan amount, loan term, CIBIL score, residential asset value, commercial asset value, luxury asset value, and bank asset value. Since the target variable has only two categories, this study treated the task as a binary classification problem.

The preprocessing step was kept close to the raw dataset. The column names were cleaned for easier coding, and the dataset was checked for missing values. No missing observations were found, so no imputation method was needed. The loan ID variable was removed because it is only an identifier and should not carry useful prediction information. The categorical fields education and self-employed were converted into numerical form, and loan status was encoded as the target label.

After preprocessing, 11 predictor variables remained. The data were divided into an 80:20 training-test split, giving 3,415 records for training and 854 records for final testing. For Logistic Regression, KNN, and SVM, the predictors were standardized the predictors before fitting the model. This scaling step is important because these algorithms are affected by feature magnitude; without it, variables measured in larger numerical units could dominate the calculation even when they are not more meaningful.

2.2 Model Selection

The five models were selected to represent different modelling ideas rather than to search for the most complex possible algorithm. Logistic Regression was used as a linear reference model. It is easy to explain and is often used in credit-related classification, so it provides a useful baseline for judging whether more flexible models add value.

Decision Tree and Random Forest were included because lending decisions often look like threshold-based rules. For instance, a credit-score threshold or a loan-term cut-off can sharply change the final decision. A Decision Tree can express such threshold rules directly. Random Forest builds many trees and averages their decisions, which usually makes the result more stable than a single tree and also produces a feature-importance ranking.

K-Nearest Neighbors was used as a distance-based method. It classifies a new applicant by looking at nearby cases in the feature space. Support Vector Machine was used as a margin-based model that searches for a boundary between approved and rejected observations. These two models were useful for comparison because they rely on assumptions different from both linear regression and tree splitting.

2.3 Evaluation Metrics

Model performance was evaluated on the test set using accuracy, precision, recall and F1-score. Accuracy measures the overall share of correctly classified records. Precision, recall, and F1-score were calculated for the approved class, because the practical concern in this dataset is whether approved applications are identified correctly without producing too many wrong approvals. Using more than one metric is necessary because a model can look strong in overall accuracy while still making costly errors for one class.

3 Results

3.1 Model Comparison

The test-set results are reported in Table 1. All five models produced usable classification results, but the gap between model families was clear. The two tree-based models ranked first and second. The Decision Tree achieved an accuracy of 0.9813 and an F1-score of 0.9849, while Random Forest reached 0.9801 in accuracy and 0.9840 in F1-score. The difference between them is very small, so the main result is that tree-based modelling fitted this dataset especially well.

Table 1. Test-set performance of the five classifiers

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.9813	0.9849	0.9849	0.9849
Random Forest	0.9801	0.9831	0.9849	0.9840
SVM	0.9450	0.9583	0.9529	0.9556
Logistic Regression	0.9133	0.9208	0.9416	0.9311
KNN	0.8946	0.9061	0.9266	0.9162

SVM was the next strongest model, with an F1-score of 0.9556. The result indicates that a margin-based classifier can separate most cases after scaling, but it still made more mistakes than the tree models. Logistic Regression and KNN were weaker. Logistic Regression obtained an F1-score of 0.9311, suggesting that a purely linear boundary was not flexible enough for the full pattern in the dataset. KNN had the lowest F1-score, 0.9162, which may be related to the mixture of financial variables and the difficulty of defining distance meaningfully across all features.

3.2 Confusion Matrix Analysis

The confusion matrix of the Decision Tree gives a clearer view of the error pattern. As shown in Table 2, the model correctly labelled 315 rejected applications and 523 approved applications. It made 16 errors in total: eight rejected applications were predicted as approved, and eight approved applications were predicted as rejected. In a lending context, the first type of error may be more risky because it could send a weak application forward, while the second type may reduce business opportunity. The balanced number of errors shows that the Decision Tree did not simply favour one class.

Table 2. Confusion matrix of the Decision Tree model

Actual / Predicted	Rejected	Approved
Rejected	315	8
Approved	8	523

Random Forest produced a very similar error level, with 17 total misclassifications. SVM, Logistic Regression, and KNN made more errors, especially when separating rejected applications from approved ones. This pattern again suggests that the data

contain threshold-like structures that are captured more naturally by tree methods than by linear or distance-based approaches.

3.3 Feature Importance Analysis

To interpret the strongest model family, the feature-importance values from Random Forest were examined. Fig. 1 shows a very uneven ranking. CIBIL score alone has an importance value of about 0.80, far higher than any other variable. Loan term is second at about 0.06, and loan amount is third at about 0.03. Income and the different asset-value variables each contribute only around 0.02, while education and self-employed status are almost zero in this model.

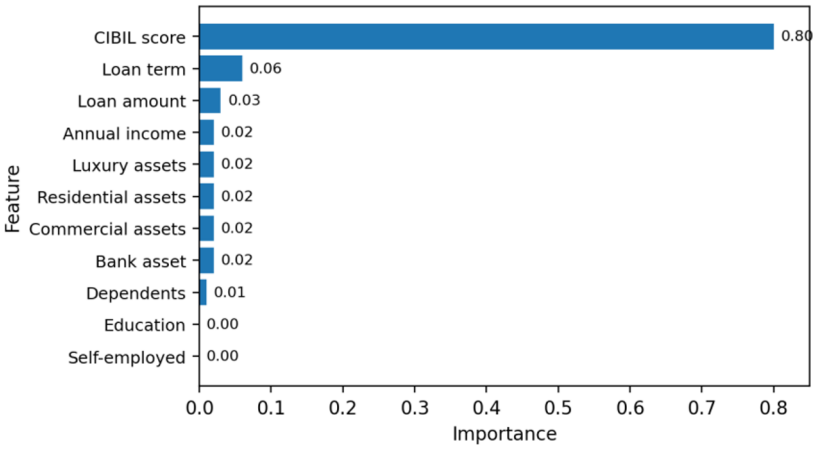


Fig. 1. Feature-importance ranking produced by the Random Forest model

This result is reasonable from a credit-risk perspective. CIBIL score directly summarizes credit history, so it is expected to carry more information than demographic or asset fields. However, the dominance of one feature also means that the model may rely heavily on credit-score thresholds. In practical use, this would require careful checking to make sure the model is not simply reproducing one rule without considering broader applicant conditions.

3.4 Discussion

The model ranking is consistent with a common finding in credit-scoring studies: models that can handle non-linear patterns often outperform a simple linear benchmark. Lessmann et al. compared many classifiers across several real credit-scoring datasets and found that advanced classifiers could exceed the industry-standard Logistic Regression approach [9]. The present result follows the same direction, because the tree models were clearly ahead of Logistic Regression on the hold-out test set.

At the same time, the fact that a single Decision Tree slightly outperformed Random Forest should not be over-interpreted. Brown and Mues found that Random

Forest and Gradient Boosting were strong on imbalanced credit-scoring datasets, while KNN and single trees could be more vulnerable under difficult class distributions [10]. In this study, the Decision Tree result may indicate that the Kaggle dataset has a relatively clean decision structure. It does not prove that a single tree is generally safer than an ensemble model. A stronger study would repeat the experiment with cross-validation, tune hyperparameters and test performance under class-imbalance adjustments.

The lower KNN result is also understandable. KNN depends on the idea that nearby records are similar, but financial variables have different scales and meanings. Even after standardization, the distance between two applicants may not represent real economic similarity very well. Zhu et al. studied Lending Club loan-default prediction and reported strong Random Forest performance after cleaning, dimensionality reduction, and SMOTE imbalance handling [11]. Compared with that work, the present experiment is simpler: it does not use SMOTE, feature selection or extensive tuning. Therefore, the result is useful as an initial comparison, but not as a final industrial model.

3.5 Limitations and Further Work

Several limitations should be noted. First, the dataset is public and relatively small, so it may not reflect the complexity of real bank data. Second, the evaluation used one train-test split. Although the test set gives an initial check, cross-validation would make the conclusion more stable. Third, the models were compared mainly with default or basic settings. Hyperparameter tuning may improve SVM, KNN or Random Forest and may change the ranking. Finally, fairness issues were not examined in this paper, such as whether errors differ across applicant groups. Future work could add cross-validation, class-imbalance treatment, fairness indicators, and SHAP-style explanations to provide a more complete assessment.

4 Conclusion

This paper compared five supervised learning models for loan approval prediction using a public dataset of 4,269 applications. After removing the identifier field, encoding categorical variables and splitting the data into training and test sets, Logistic Regression, Decision Tree, Random Forest, KNN and SVM were evaluated with accuracy, precision, recall and F1-score. The strongest results came from the tree-based models. Decision Tree achieved the highest test-set performance in this experiment, while Random Forest was nearly equal and offered a useful feature-importance view.

The most important predictor in the Random Forest output was CIBIL score, followed by loan term and loan amount. This suggests that creditworthiness and repayment-related variables were central to the prediction result. Overall, the experiment shows that machine learning can be a useful support tool for structured loan-approval screening. However, the conclusion should remain cautious: a real

lending system would need stronger validation, interpretability checks, fairness testing and domain review before being used in practice.

References

1. James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J.: An Introduction to Statistical Learning: With Applications in Python. 3rd edn. Springer (2023)
2. Khandani, A.E., Kim, A.J., Lo, A.W.: Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance 34(11), 2767–2787 (2010)
3. Dastile, X., Celik, T., Potsane, M.: Statistical and machine learning models in credit scoring: A systematic literature survey. Applied Soft Computing 91, 106263 (2020)
4. Dumitrescu, E., Hue, S., Hurlin, C., Tokpavi, S.: Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research 297(3), 1178–1192 (2022)
5. de Lange, P.E., Melsom, B., Vennerod, C.B., Westgaard, S.: Explainable AI for credit assessment in banks. Journal of Risk and Financial Management 15(12), 556 (2022)
6. Liu, Y., Zhang, C., Matousek, R.: Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. Research in International Business and Finance 59, 101567 (2022)
7. Guo, W., Zhou, Z., Qian, C.: Application of mobile app list information in credit risk evaluation: An empirical study based on an online lending platform. Chinese Journal of Management Science 30(12), 96–107 (2022)
8. Sharma, A.: Loan-Approval-Prediction-Dataset, https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset?resource=download, last accessed 2026/06/17
9. Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 247(1), 124–136 (2015)
10. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications 39(3), 3446–3453 (2012)
11. Zhu, L., Qiu, D., Ergu, D., Ying, C., Liu, K.: A study on predicting loan default based on the random forest algorithm. Procedia Computer Science 162, 503–513 (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

