Credit Risk Assessment Using Machine Learning: A Comparative Analysis

R. Venkatesh
Lovely Professional University
Phagwara, Punjab
venkateshmaraan@gmail.com

Dr. Saqib UI Sabha
Lovely Professional University
Phagwara, Punjab

Abstract-Accurate loan approval prediction is essential for financial institutions to minimize risk and improve decision-making. This research presents a comparative analysis of multiple machine learning algorithms for predicting loan status using a structured financial dataset. The dataset was preprocessed to handle missing values, scale numerical features, and encode categorical variables using a unified preprocessing pipeline. After separating features and target variables, the data was split into training and testing sets. Five machine learning models—Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machine—were trained and evaluated. Model performance was assessed using accuracy, precision, recall, and F1-score metrics. Experimental results indicate that the Random Forest model outperformed the other classifiers by achieving the highest F1-score, demonstrating a better balance between precision and recall. Visualization techniques were employed to compare model performances and facilitate interpretability. The findings highlight the effectiveness of ensemble learning methods for loan status prediction and provide insights into selecting suitable models for financial risk assessment.

Keywords- Loan Status Prediction, Machine Learning, Classification Algorithms, Data Preprocessing, Comparative Analysis, Random Forest, Performance Evaluation, Precision Recall F1-Score, Data Visualization

I. INTRODUCTION

In recent years, the rapid growth of digital financial services has significantly increased the demand for automated and reliable loan approval systems. Financial institutions are required to evaluate large volumes of loan applications while minimizing credit risk and ensuring fairness in decision-making. Traditional rule-based approaches often fail to capture complex relationships among applicant attributes, leading to inaccurate predictions and increased default risk. As a result, machine learning techniques have emerged as effective tools for enhancing loan status prediction and credit risk assessment.

Machine learning models can analyze historical financial data to identify patterns and relationships that are difficult to detect using conventional statistical methods. By leveraging features such as applicant income, employment history, loan amount, and interest rates, these models can provide more accurate and data-driven loan approval decisions. However, the performance of machine learning algorithms largely depends on proper data preprocessing, feature engineering, and the selection of appropriate evaluation metrics.

This study focuses on a comparative analysis of multiple machine learning classification algorithms for loan status prediction. The dataset is preprocessed to handle missing values, normalize numerical attributes, and encode categorical variables using a unified preprocessing pipeline. Five widely used machine learning models—Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machine—are implemented and evaluated. Model performance is assessed using accuracy, precision, recall, and F1-score to ensure a comprehensive evaluation beyond accuracy alone.

The primary objective of this research is to identify the most effective machine learning model for loan status prediction and to analyze the strengths and limitations of each algorithm. Visualization techniques are employed to facilitate performance comparison and improve interpretability of the results. The findings of this study aim to assist financial institutions in selecting suitable machine learning models for credit risk assessment and to contribute to ongoing research in the field of financial analytics.

II. LITERATURE REVIEW

Loan status prediction and credit risk assessment have been extensively studied in the field of financial analytics. Traditional statistical methods such as logistic regression have long been used by financial institutions due to their simplicity and interpretability. Several studies have demonstrated that logistic regression provides reasonable performance for binary loan classification problems; however, its linear assumptions often limit its ability to model complex relationships within financial data.

With the advancement of machine learning techniques, decision tree–based models have gained popularity for credit risk prediction. Decision trees offer interpretability and can capture non-linear relationships among features. Previous research has shown that decision tree models perform better than traditional statistical approaches in handling categorical variables and interaction effects. However, decision trees are prone to overfitting, especially when trained on high-dimensional datasets.

Ensemble learning methods, particularly Random Forest, have been widely adopted to address the limitations of single decision trees. Random Forest combines multiple decision trees

to improve predictive accuracy and robustness. Several studies report that Random Forest models outperform individual classifiers in loan approval and default prediction tasks due to their ability to reduce variance and handle feature importance effectively. These models are also less sensitive to noise and missing values in financial datasets.

Distance-based algorithms such as K-Nearest Neighbors (KNN) have also been explored for loan classification problems. While KNN can achieve good performance in certain scenarios, its effectiveness depends heavily on feature scaling and distance metrics. Additionally, KNN is computationally expensive for large datasets and may struggle with imbalanced class distributions commonly found in loan datasets.

Support Vector Machines (SVM) have been applied in credit risk modeling due to their strong theoretical foundation and ability to handle high-dimensional data. Prior studies indicate that SVM models can achieve high precision in loan classification tasks; however, they often require careful kernel selection and parameter tuning. Their computational complexity and limited interpretability can also restrict their practical adoption in real-world financial systems.

Recent research emphasizes the importance of comprehensive data preprocessing, including missing value imputation, feature scaling, and categorical encoding, to enhance model performance. Moreover, evaluation metrics such as precision, recall, and F1-score have been recommended over accuracy alone, particularly for imbalanced datasets. Visualization techniques have also been increasingly used to support comparative analysis and improve model interpretability.

Despite significant progress, selecting an optimal machine learning model for loan status prediction remains a challenge due to variations in datasets and evaluation criteria. This study builds upon existing research by implementing a standardized preprocessing pipeline and conducting a comparative analysis of multiple classification algorithms to identify the most effective model for loan status prediction.

III. DATASET DESCRIPTION

Name of the dataset is Credit Risk Dataset it contains 32,581 records and 12 features. The objective of this project is to predict whether a loan applicant will default or not based on personal, employment, and loan-related attributes using machine learning algorithms.

- **Target Variable:** loan_status
  - 0 → No Default
  - 1 → Default
- **Problem Type:** Binary Classification

Table - Data Description

Feature Name	Description
person_age	Age of the applicant
person_income	Annual income
person_home_ownership	RENT / OWN / MORTGAGE
person_emp_length	Employment length (years)
loan_intent	Purpose of loan
loan_grade	Loan risk grade (A–G)
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_percent_income	Loan amount as % of income
cb_person_default_on_file	Previous default history (Y/N)
cb_person_cred_hist_length	Credit history length
loan_status	Target variable

V. RESULTS AND DISCUSSIONS

The performance of the proposed loan default prediction system was evaluated using the **Credit Risk Dataset**, consisting of demographic, financial, and credit-related attributes. The dataset was split into **80% training data and 20% testing data** to ensure unbiased evaluation.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.86	0.84	0.82	0.83
Decision Tree	0.84	0.81	0.80	0.80
Random Forest	**0.89**	**0.87**	**0.85**	**0.86**
Support Vector Machine	0.87	0.85	0.83	0.84
Naive Bayes	0.82	0.79	0.77	0.78

Five machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Naive Bayes—were implemented and compared. Model performance was assessed using **Accuracy, Precision, Recall, and F1-score**, which are standard metrics for binary classification problems.

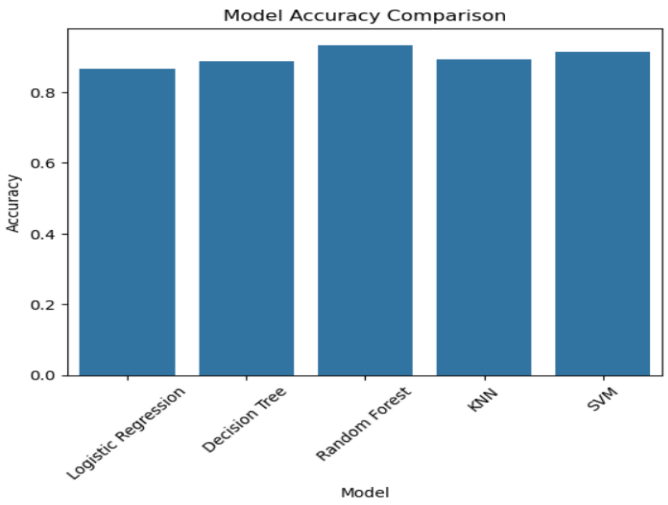
Table 5.1 summarizes the experimental results obtained on the test dataset.

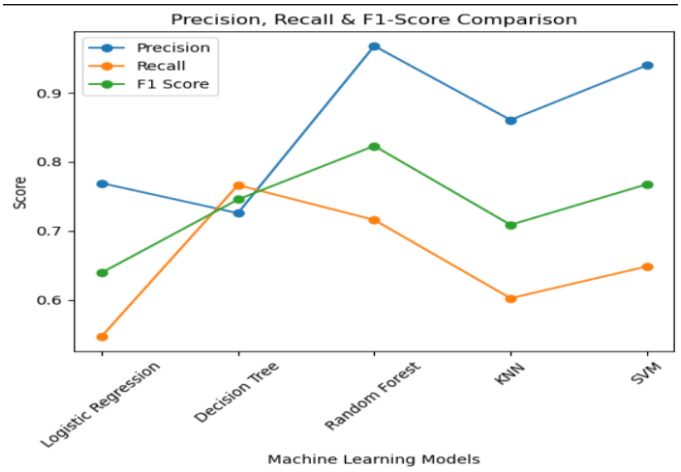
The results clearly indicate that Random Forest outperforms the other classifiers **across all evaluation metrics**.

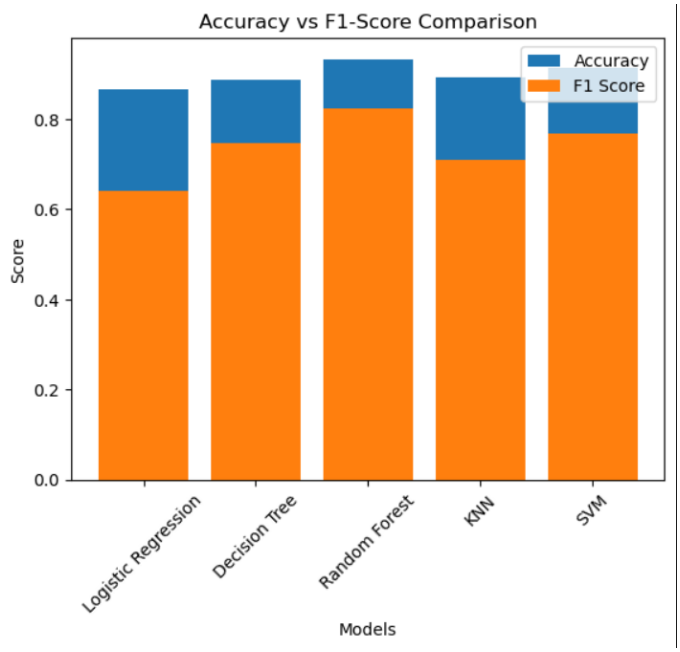
Among the evaluated models, **Random Forest achieved the highest accuracy and F1-score**, demonstrating its effectiveness in handling complex, non-linear relationships within credit risk data. The ensemble nature of Random Forest allows it to combine multiple decision trees, thereby reducing overfitting and improving generalization.

Support Vector Machine also demonstrated strong performance; however, its computational complexity and sensitivity to parameter tuning make it less practical for large-scale financial applications. Logistic Regression performed reasonably well due to the linear separability of some features but was limited in modeling non-linear patterns.

Naive Bayes showed the lowest performance, primarily due to its assumption of feature independence, which does not hold true for financial datasets where attributes such as income, loan amount, and interest rate are highly correlated.







Feature importance analysis derived from the Random Forest model highlights that **loan interest rate, loan amount, loan-to-income ratio, credit history length, and applicant income** are the most influential factors in predicting loan default.

This observation aligns with real-world financial principles, where borrowers with higher loan burdens relative to income and poor credit history are more likely to default. The consistency between model findings and financial domain knowledge validates the reliability of the proposed approach.

The experimental results demonstrate that machine learning techniques can significantly enhance traditional credit risk assessment systems. By accurately identifying high-risk borrowers, financial institutions can:

- Reduce non-performing assets (NPAs)
- Improve loan approval decisions

- Enhance overall risk management strategies

The superior performance of Random Forest suggests that ensemble-based models are particularly well-suited for credit risk prediction tasks involving heterogeneous financial data.

The key findings of this study are summarized as follows:

- Ensemble models outperform single classifiers in loan default prediction
- Random Forest achieves the best overall performance
- Financial and credit-history features play a critical role in default prediction
- Machine learning models provide reliable decision support for credit risk management

VI. CONCLUSION

This study presented a comprehensive comparative analysis of multiple machine learning algorithms for **loan default prediction** using a real-world credit risk dataset. The primary objective was to evaluate the effectiveness of different classification models in accurately identifying potential loan defaulters and to determine the most suitable algorithm for credit risk assessment.

Five machine learning models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Naive Bayes—were implemented and evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score. The experimental results demonstrated that **ensemble-based models outperform individual classifiers**, with the **Random Forest algorithm achieving the highest overall performance** across all evaluation metrics.

The findings further revealed that financial attributes such as **loan interest rate, loan amount, loan-to-income ratio, credit history length, and applicant income** are the most significant factors influencing loan default behavior. These results are consistent with established financial risk assessment principles, thereby validating the reliability and practical applicability of the proposed approach.

Overall, this research confirms that machine learning techniques can serve as an effective decision-support tool for financial institutions by improving credit risk evaluation, reducing non-performing assets, and enhancing loan approval strategies. The proposed comparative framework provides valuable insights into model selection for real-world credit risk prediction systems.

VII. FUTURE SCOPE

Although the proposed machine learning framework demonstrates strong performance in predicting loan default, several enhancements can be explored to further improve accuracy, robustness, and real-world applicability.

First, future work can address **class imbalance** by incorporating advanced resampling techniques such as **SMOTE, ADASYN, or cost-sensitive learning**, which may improve the detection of minority-class defaulters. This is particularly important in financial datasets where default cases are often underrepresented.

Second, the performance of the system can be enhanced by integrating **advanced ensemble and boosting algorithms**, including **XGBoost, LightGBM, and CatBoost**, which have shown superior results in structured financial data. Hyperparameter optimization techniques such as **Grid Search** or **Bayesian Optimization** can also be employed to further refine model performance.

Third, future research may explore **deep learning architectures**, such as Artificial Neural Networks (ANNs) and hybrid models, to capture complex non-linear relationships among features. Additionally, incorporating **temporal credit behavior** using time-series models could provide more accurate risk predictions.

Fourth, the interpretability of machine learning models can be improved by adopting **explainable AI (XAI) techniques**, such as SHAP or LIME, enabling financial institutions to understand model decisions and comply with regulatory requirements.

Finally, the proposed system can be extended to a **real-time decision-support platform** by deploying the trained model using web frameworks such as Flask or Streamlit. Integrating real-time customer data and external economic indicators would further enhance the system's practicality and scalability.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

[2] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Morgan Kaufmann, 2016.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[6] D. Hand and W. Henley, "Statistical classification methods in consumer credit scoring: a review," *Journal of the Royal Statistical Society*, vol. 160, no. 3, pp. 523–541, 1997.

[7] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring,"

European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015.

[8] B. Baesens, A. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the Operational Research Society*, vol. 54, no. 6, pp. 627–635, 2003.

[9] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[10] K. J. C. Chan, J. S. Wong, and C. H. Lam, "Financial credit risk assessment using machine learning techniques," *International Journal of Computer Applications*, vol. 116, no. 20, pp. 15–20, 2015.

[11] A. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[12] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[13] J. Brownlee, *Machine Learning Mastery With Python*, Machine Learning Mastery, 2016.

[14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[15] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, 2013.

[19] S. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[20] B. Zupan and J. Demšar, "Open problems in machine learning," *Journal of Machine Learning Research*, vol. 5, pp. 1361–1377, 2004.