International Journal of Science, Strategic Management and Technology
Volume **02** Issue **11** November-2026 | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Predictive Analytics for Credit Risk Assessment: A Machine Learning Approach to Loan Default Prediction

Dr. Kalyani Gorti ,A.Anuraag , Eguri Helen Senorita
Assistant Professor, ,Students, Bcom Hons(Business Analytics)
Department of Commerce
Bhavan's Vivekananda College of Science, Humanities and Commerce,
Sainikpuri, Secunderabad

https://doi.org/10.55041/ijsmt.v2i11.014

Cite this Article: A.Anuraag, & Senorita, E. H. (2026). Predictive Analytics for Credit Risk Assessment: A Machine Learning Approach to Loan Default Prediction. International Journal of Science, Strategic Management and Technology, *02*(11), 1-9. https://doi.org/10.55041/ijsmt.v2i11.014

License: This article is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting use, distribution, and reproduction in any medium, provided the original author(s) and source are properly credited.

Abstract

The integration of Machine Learning (ML) into financial decision-making helps the financial to strengthen credit-risk assessment and improve lending decisions. Traditional credit assessment methods that depend on fixed criteria and historical records, often struggle to capture complex relationships among borrower and loan characteristics. Machine learning helps bridge that gap. This study examined the application of Machine Learning techniques to loan default prediction. Using a dataset comprising 45,000 loan records, the study analysed borrower demographic, financial, employment, credit-history and loan-related characteristics to identify factors associated with default risk. The study applies three supervised machine learning classification models namely Logistic Regression, Decision Tree, and Random Forest and compared results using accuracy, precision, recall, F1-score and ROC-AUC. Random Forest performed best overall (92.69% accuracy, 88.19% precision, 81.51% F1-score, 97.04% ROC-AUC), while Decision Tree achieved the highest recall (77.46% vs. 75.77%), making it slightly better at identifying actual defaulters. Logistic Regression also performed well, reaching 94.95% ROC-AUC. Random Forest offers the strongest overall performance, but higher-recall models are valuable when minimizing missed high-risk borrowers is the priority. Overall, Machine Learning proved a practical tool for strengthening credit-risk assessment and supporting better financial decisions.

Keywords: *Predictive Analytics, Machine Learning, Credit Risk Assessment, Loan Default Prediction, Classification Models.*

Introduction

The rapid growth of digital financial services and data-driven technologies has transformed the way financial institutions assess risk and make lending decisions. Traditional credit assessment methods often rely on predefined criteria and historical financial information, while the increasing availability of large and diverse datasets has created opportunities to develop more dynamic and predictive approaches. In this context, predictive analytics and machine learning have emerged as important technology-enabled tools for supporting financial decision-making.

Loan default is a significant concern for lending institutions because unexpected defaults can result in financial losses, increased credit risk, and inefficient allocation of financial resources. Identifying borrowers who are more likely to default before a lending decision is made can therefore help institutions strengthen their risk-management practices. Factors such as income, credit score, loan amount, interest rate, employment experience, loan-to-income ratio, and previous default history may provide valuable information for assessing a borrower's credit risk.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

Machine learning provides an opportunity to analyse these factors collectively and identify patterns that may not be easily captured through conventional assessment methods. By using historical loan data, predictive models can estimate the likelihood of default and provide decision-support information to financial institutions. This shifts credit assessment from a predominantly reactive process toward a more data-driven and proactive approach.

This study therefore examines the application of machine learning for loan default prediction as an example of Predictive Analytics for Financial Decision-Making under the broader theme of Technology-Enabled Innovations. The study compares three machine learning techniques - Logistic Regression, Decision Tree, and Random Forestto determine their effectiveness in predicting loan default. In addition to comparing model performance, the study identifies important factors associated with default risk. The findings aim to demonstrate how predictive analytics can support credit-risk assessment and enable more informed, efficient, and proactive lending decisions.

Review of Literature

Li and Wu (2024) compared nine machine-learning models for loan default prediction and found Random Forest to be the most efficient and stable. Using SHAP, the study identified key factors such as age and work experience influencing default risk, combining predictive accuracy with model interpretability.Overall, the paper combined accurate loan-default prediction with explainable AI, making the results more understandable for financial institutions.

Lin (2024) compared four machine-learning modelsi.eLogistic Regression, Random Forest, XGBoost, and AdaBoostfor loan default prediction. XGBoost performed best, while Logistic Regression performed relatively poorly, highlighting the potential of ML, particularly XGBoost, to improve credit-risk assessment and reduce lending losses.

Zhang et al. (2025) compared Gradient Boosting, XGBoost, LightGBM, and Random Forest for loan default prediction within an automated BPM framework. Gradient Boosting achieved the highest accuracy (88.87%), while XGBoost had the highest ROC-AUC (97.14%). Key predictors included credit score, debt-to-income ratio, interest rate, and previous defaults, though the study was limited by synthetic data, limited interpretability, and insufficient fairness testing.

Xu et al. (2023) integrated Explainable AI (XAI) into loan default prediction to balance model accuracy and transparency. The study used global and local explanations to identify key risk factors and support human review, improving the interpretability of automated credit decisions. However, it was limited by correlational feature attribution, potential dataset bias, and reliance on post-hoc explainability rather than comprehensive fair-ML auditing.

SharafEldin, M.A. (2025)analyzed agricultural credit data using Decision Tree, Random Forest, KNN, and Gradient Boosting models to predict loan defaults. The Decision Tree performed best, achieving 88% accuracy, 53% recall, and 89% specificity, while offering simple interpretation for banking staff. However, class imbalance, inconsistent metrics, unclear data descriptions, and limited validation details reduce the study's reliability and reproducibility.

Mor et al. (2022) examined the application of artificial intelligence in predicting loan-default risk in Indian commercial banks. Using a supervised machine learning approach based on logistic regression, the study classified banks according to their level of loan risk. The findings indicated that interest income and spread were negatively associated with loan-default risk, while interest expenditure, operating expenses and the credit-deposit ratio showed a positive association. The study highlights the potential of machine learning techniques in supporting loan-risk assessment in the Indian banking sector.

Need for the Study

Managing credit risk has become critical for the financial industry since the inability of the borrowers to repay the loans can impact the performance of the lenders. Credit analysis has been traditionally carried out using credit scores, repayment history, income level, and other predetermined variables. However, these techniques might not always incorporate all the variables involved in the decision-making process and leading to default by the borrower. With increasing availability of information on the finances of customers, it has become important to find better techniques for analysing such information and spotting at-risk borrowers. Predictive modelling through machine learning is one

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

way of doing so since it involves the examination of historical data for prediction of the probability of loan default. The current study, thus, becomes necessary for finding the role of such techniques in credit-risk assessment.

Objectives of the Study

1. To examine the role of predictive analytics in supporting credit-risk assessment and financial decision-making.
2. To identify key borrower and loan-related factors associated with loan default.
3. To compare the effectiveness of different classification models in predicting credit risk.
4. To evaluate the predictive performance of the models using accuracy, precision, recall, F1-score, and ROC-AUC.

Scope of the Study

- Dataset Scope - The study uses a dataset comprising 45,000 loan records, covering borrower demographic, financial, employment, credit-history, and loan-related characteristics.
- Analytical Scope - The study focuses on predicting loan default and identifying the borrower and loan characteristics associated with default risk using predictive analytics.
- Conceptual Scope - The study focuses on the role of technology-enabled predictive analytics in financial decision-making, rather than examining broader aspects of banking operations or loan recovery

Limitations of the Study

1. Dataset Limitation - The findings may not fully represent borrowers across different financial institutions, regions, or lending markets.
2. Variable Limitation - The analysis is restricted to the demographic, financial, employment, credit-history, and loan-related variables available in the dataset. Other factors that may influence loan default, such as changes in economic conditions or borrower behaviour over time, are not considered.
3. Class Imbalance- The dataset exhibited a class imbalance (78% were non-defaulters to 22% defaulters), but no resampling techniques were employed to fix this issue and hence could have contributed to the relatively low recalls recorded in the three classifiers

Research Methodology

This study adopted a quantitative research approach to examine the application of predictive analytics in loan default prediction and credit-risk assessment.

Data Source:

The study is based on secondary data obtained from a publicly available loan dataset on Kaggle. The dataset consists of 45,000 loan records containing information related to borrowers' demographic characteristics, financial position, employment, credit history, and loan attributes.

The dataset was first organized and examined for missing values, duplicate observations, inconsistent data types, and other data-quality issues. Categorical variables were converted into a suitable numerical format for machine learning analysis, while numerical variables were prepared for model development.

Variables Used:

The dataset used in the analysis contains borrower and loan-related variablessuch as age, income, employment experience, home ownership, loan amount, loan intent, interest rate, loan-to-incomeratio, credit score, and previous loan defaults.
Loan status was the dependent variable, indicating whether the borrower defaulted or not.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **11** November-2026 | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Data Analysis and Machine Learning Techniques:

The analysis begins with descriptive and exploratory data analysis to understand the distribution of borrower and loan characteristics and identify patterns between the variables and loan-default outcomes. Graphical techniques are used to examine relationships and differences between borrowers who defaulted and those who did not.

Following the exploratory analysis, three machine learning classification models ie logistic regression, decision tree and random forest are developed to predict loan default. The models are trained using the prepared dataset, which is divided into 80% training data (30,039 observations) and 20% testing data (7,510 observations) using a stratified train-test split. The training set is used to build the predictive models, while the testing set contains unseen observations used to assess their predictive performance. The comparative approach allows the study to determine which model provides more effective support for identifying potential loan defaults.

Model Evaluation:

The performance of the models is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy measures the overall proportion of correctly classified observations, while precision indicates how accurately the model identifies predicted defaulters. Recall measures the model's ability to correctly identify borrowers who actually defaulted. F1-score provides a balance between precision and recall, while ROC-AUC measures the model's ability to distinguish between default and non-default cases.

Feature Importance Analysis:

The study also examines feature importance to identify the borrower and loan-related characteristics that contribute most significantly to loan-default prediction. This analysis helps move beyond simply determining which model performs best and provides insights that can be useful for credit-risk assessment and financial decision-making.

Data Analysis Tools:

The data preparation, analysis, visualization, and machine learning modelling were performed using Python in Google Colab. Python libraries were used for data manipulation, exploratory data analysis, visualization, model development, and evaluation. Microsoft Excel was used for preliminary data organization and preparation of the dataset before its analysis in Python.

Analysis & Interpretations

1. Loan Status Distribution:The analysis of loan status distribution showed that the majority of borrowers did not default on their loans, with approximately 29,500 non-default cases compared to around 8,000 default cases. This indicates that most borrowers were able to meet their repayment obligations, while a smaller proportion experienced loan default. Although defaults represented a minority of the observations, they constituted a significant source of credit risk for lending institutions.

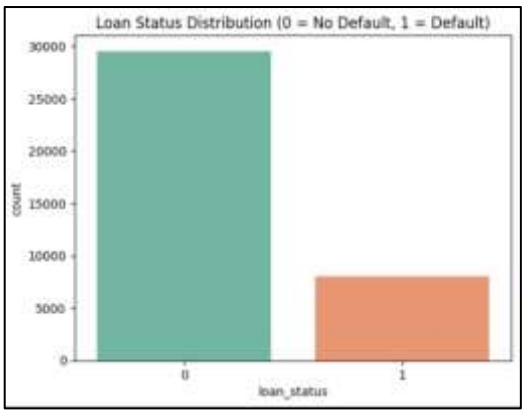


Figure 2/Loan Status by Loan Purpose

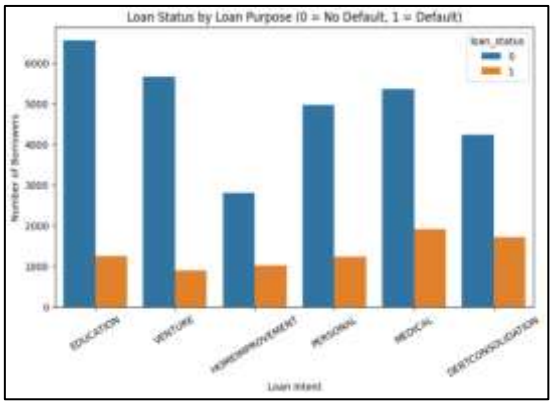


Figure 1/Loan Status Distribution

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **11** November-2026 | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

2. Loan Status by Loan Purpose: Medical and debt-consolidation loans recorded relatively higher numbers of defaults compared to other loan purposes, while venture loans exhibited lower default levels. This suggests that the purpose for which a loan is obtained may influence repayment behaviour and default risk. Borrowers seeking medical loans may face unexpected financial burdens, whereas those applying for debt-consolidation loans may already be experiencing financial stress. Therefore, loan purpose can serve as a useful indicator in assessing borrower risk and improving credit-risk evaluation.
3. Loan Status by Previous Loan Default History: The analysis of previous loan default history revealed a strong relationship between past and future repayment behaviour. Borrowers with a history of previous loan defaults exhibited a substantially higher likelihood of defaulting again compared to borrowers with no prior default history. Therefore, previous default history can serve as a valuable predictor in credit-risk assessment and loan default prediction models.

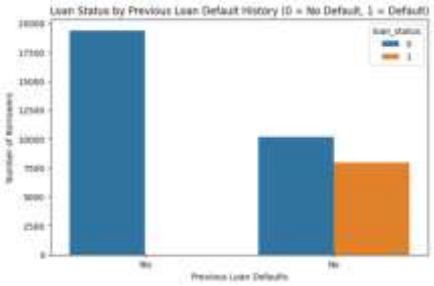


Figure 3/Loan Status by Previous Loan Default History

4.Model Performance Evaluation

	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC-AUC (%)
0	Logistic Regression	89.24	75.73	72.70	74.19	94.95
1	Decision Tree	89.63	74.70	77.46	76.05	85.19
2	Random Forest	92.69	88.19	75.77	81.51	97.04

Figure 4/Model Performance Evaluation

The model evaluation showed that all three models achieved relatively strong predictive performance, with Random Forest providing the best overall results. It recorded the highest accuracy (92.69%), precision (88.19%), F1-score (81.51%), and ROC-AUC (97.04%), indicating strong overall classification and discrimination between defaulters and non-defaulters. Decision Tree achieved the highest recall (77.46%), suggesting that it was the most effective at identifying actual defaulters. Logistic Regression achieved 89.24% accuracy and 94.95% ROC-AUC, demonstrating good performance while offering greater interpretability. Overall, the results indicated that Random Forest was the strongest overall model, while Decision Tree could be more suitable when identifying as many potential defaulters as possible is the primary objective.

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

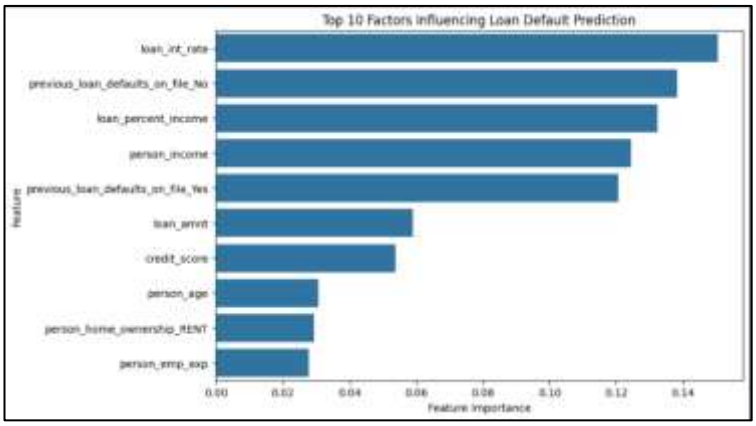


Figure 5/Feature Importance Analysis

Feature importance analysis was performed using the Random Forest model to identify the variables that contributed most significantly to loan default prediction. The results revealed that interest rate was the most influential predictor, followed by previous default history, loan-to-income ratio, income, and loan amount. In contrast, demographic factors such as age, employment experience, and home ownership status exhibited relatively lower importance. These findings suggest that financial and credit-related characteristics play a greater role in determining default risk than personal or demographic attributes. Therefore, lenders can improve credit-risk assessment by placing greater emphasis on key financial indicators when evaluating loan applications.

Findings of the Study

The study demonstrated that machine learning techniques can be effectively applied to loan default prediction and credit-risk assessment. All three classification models achieved strong predictive performance, with accuracy exceeding 89%. Among them, Random Forest emerged as the best-performing model, achieving the highest accuracy (92.69%), precision (88.19%), F1-score (81.51%), and ROC-AUC (97.04%). These results indicate its strong ability to distinguish between defaulters and non-defaulters. However, the Decision Tree achieved the highest recall (77.46%), making it more effective in identifying actual defaulters and reducing the risk of overlooking high-risk borrowers.

Exploratory data analysis revealed important patterns in loan default behaviour. Most borrowers successfully repaid their loans, while defaults represented a smaller proportion but remained a significant credit-risk concern. Loan purpose also appeared to influence repayment behaviour, with medical and debt-consolidation loans showing relatively higher default levels compared to other categories.

The study further identified a strong relationship between previous default history and future default risk. Borrowers with a history of default were considerably more likely to default again, highlighting the importance of repayment history in credit evaluation. Feature importance analysis using Random Forest identified interest rate as the most influential predictor, followed by previous default history, loan-to-income ratio, income, and loan amount. These findings suggest that financial and credit-related variables have a greaterinfluence on default prediction than demographic factors such as age and employment experience.

Overall, the findings indicate that no single model is optimal for every lending objective. While Random Forest provided the strongest overall predictive performance, models with higher recall may be preferable when the primary goal is to identify as many potential defaulters as possible. Thus, machine learning can serve as a valuable technology-enabled tool for strengthening credit-risk management, early risk identification, and informed lending decisions.

Conclusion

This study clearly shows how machine learning algorithms can support credit risk assessment since all three models show high level of prediction on borrowers' and loans' data. Random forest turned out to be the most accurate and powerful algorithm when high recall score is required. Decision Tree should be used since it is more useful for

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.

catching actual defaulters, something that should be considered instead of blindly choosing the algorithm with the highest accuracy.

Interest rate and previous default history, which are related to finances of the borrower, were the most important factors for predicting default, while demographic information had low impact. These results confirm that lenders' decisions should be based on the borrower's financial behaviour rather than personal information. Overall, the results show how predictive analytics, with the help of modern technology, can be effectively used to improve credit risk assessment process.

References

- Li, H., & Wu, W. (2024). Loan default predictability with explainable machine learning. https://doi.org/10.1016/j.frl.2023.104867
- Lin, J. (2024). Research on loan default prediction based on logistic regression, random forest, XGBoost and AdaBoost. https://doi.org/10.1051/shsconf/202418102008
- SharafEldin, M. A. (2025). A proposed framework for loan default prediction using machine learning techniques. https://thesai.org/Downloads/Volume16No6/Paper_40A_Proposed_Framework_for_Loan_Default_Prediction.pdf
- Zhang, X., & Zhang, T. (2025). Data-driven loan default prediction: A machine learning approach for enhancing business process management. https://doi.org/10.3390/systems13070581
- Zhu, X., & Chu, Q. (2023). Explainable prediction of loan default based on machine learning models. https://doi.org/10.1016/j.dsm.2023.04.003
- Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2021). Machine learning in finance: A topic modeling approach. *European Financial Management.*, **28**, 744–770. **https://doi.org/10.1111/EUFM.1232**
- Mor, S., Aneja, R., Madan, S., &Gupta, S. (2022). *Artificial intelligence and loan default: The case of commercial banks in India*. *Strategic Change*, 31(6), 571–580. https://doi.org/10.1002/jsc.2529
- N. Kumar, "The Impact of Artificial Intelligence on Financial Decision-Making," Journal of Information Systems Engineering & Management, vol. 10, p. 473, May 2025, doi: 10.52783/jisem.v10i47s.9307. CrossRef , Google Scholar
- Dr. K. Naveen Kumar, From Adoption to Advantage: How Generative AI Capability Drives Integrated Business Performance in Emerging Markets. (2026). MSW Management Journal, 36(2), 647-650.
- Dr. K. Naveen Kumar, Dr. Darshan Subherwal, Ms. R. Renuka, Dr. Kavitha Rapolu. (2026). Employee Productivity and Digital Transformation in Modern Organizations . Journal of Daoist Studies, 19(S2), 1407–1414. Retrieved from https://journalofdaoiststudies.org/index.php/journal/article/view/394
- Kumar, K.N.; Upadhyay, D.M.; Maheshwari, P.D.D. An Examination of Financial Performance Before and after Mergers: A Case Study of Selected Indian Pharmaceutical Companies. J. Res. Adm. 2023, 5, 4132–4142. [Google Scholar]

© Author(s). This work is peer-reviewed, openly published, and permanently archived
This article is openly accessible and reusable with proper attribution.