2026 Volume: 02 Issue: 08 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-09

CREDIT RISK ASSESSMENT MODEL FOR INDIAN BANK USING MACHINE LEARNING

R. Vineeth Swaroop*[1] and Dr R. N. Kulkarni[2]

[1]Department of Management Studies, Ballari Institute of Technology and Management, Ballari.

[2]Department of CSE, Ballari Institute of Technology and Management, Ballari.

Article Received: 16 July 2026
Article Revised: 05 August 2026
Published on: 25 August 2026

***Corresponding Author: R. Vineeth Swaroop**
Department of Management Studies, Ballari Institute of Technology and Management, Ballari.

DOI: https://doi-doi.org/101555/ijrpa.1667

ABSTRACT

In banking, proper risk assessment is an essential task to gauge the probability of getting back the loan. In the case where such an assessment fails on a wide scale, the results manifest themselves in the form of non-performing assets, which lead to a loss in profits for banks and also a destabilization of the overall financial sector. Credit risk assessment involves manually checking documents and applying some scoring criteria. In the modern world, when banks accumulate huge amounts of data regarding their clients, traditional methods cannot be applied efficiently. Therefore, it is necessary to investigate whether supervised machine learning techniques are able to offer a better alternative. This study, "Credit Risk Assessment Model for Indian Bank Using Machine Learning," employs a secondary dataset containing 51,336 observations with 87 features related to client demographics, credit score history, loan account details, repayment history, credit inquiries, usage ratio, employment and bureau credit score. Prior to model construction, the data was preprocessed by removing missing values, label-encoding categorical columns, and selecting the best predictors using correlation and mutual information analysis. Four classification models were developed and compared: Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). Evaluation criteria included accuracy, precision, recall, F1-score, confusion matrix performance, and 5-fold cross-validation. While all four models achieved excellent accuracy on the four-class objective (Approved_Flag: P1–P4), tree-based ensemble techniques clearly surpassed the linear model. The accuracy of Logistic Regression is 96.65%, Decision Tree 99.15%, Random Forest 99.41%, and XGBoost 99.48%.

INDEX TERMS: Credit risk, Machine learning, Random Forest, Extreme Gradient Boosting (XGBoost).

I. INTRODUCTION

The banking sector makes an essential contribution to the economic development of a country through the provision of deposits, loans, investments and credit facilities. Among these services, lending is the primary source of income for banks. However, such operations carry credit risk, i.e., the probability of default in loan repayment by customers. Poor credit risk management leads to loan default, an increased number of nonperforming assets (NPAs), lower profitability and financial instability. Traditional appraisal methods, though effective, are subjective, time-consuming and struggle to scale with the growing volume of digital financial data.

Modern developments in artificial intelligence, machine learning and data analytics offer new opportunities for processing large volumes of financial data, uncovering hidden patterns and predicting loan default more efficiently than conventional approaches. This project, "Credit Risk Assessment Model for Indian Bank Using Machine Learning," builds a predictive system that classifies loan applicants according to their creditworthiness. The study uses a secondary Indian credit dataset comprising 51,336 customer records and 87 attributes covering demographics, credit history, repayment behaviour, loan details and financial exposure. Four supervised algorithms — Logistic Regression, Decision Tree, Random Forest and XGBoost — are implemented and compared using Accuracy, Precision, Recall, F1-score, Confusion Matrix and Cross-validation.

The problem addressed is to develop a machine learningbased credit risk assessment model for Indian Bank that accurately predicts the creditworthiness of loan applicants, enabling faster, more reliable and data-driven lending decisions while minimizing the risk of NPAs. The objectives are to analyse credit risk factors from internal bank and external CIBIL data, preprocess the data for model development, implement and compare the four algorithms, and evaluate them using standard classification metrics.

II. LITERATURE REVIEW

A number of recent studies have examined credit risk in the Indian banking context using both statistical and machine learning approaches. Prabha et al. used structural equation modelling on survey data from 213 banking professionals and found that financial health analysis has a significant positive impact on NPAs, while credit history alone was not significant. Kumar and Srinivas compared Logistic Regression, Decision Trees, Random Forest, Neural Networks and SVM for provisioning under Basel III/IFRS 9, reporting that Random Forest achieved the highest accuracy (92%) and precision (88%).

Arora and Kumar proposed a composite Credit Risk Management (CRM) index for 35 scheduled commercial banks, obtaining an average score of 64.66%. Koju et al. used dynamic panel GMM estimation on 50 Indian banks (2000– 2013) and found that capital adequacy and bank size positively influence NPAs, while profitability and income diversification reduce credit risk. Arora further examined the effect of branchnetwork size on CRM practices, finding no statistically significant difference across network sizes.

Kumar and Gupta applied the Altman Z-Score model to five Indian capital-goods firms, while Das and Deb studied the relationship between Capital Adequacy Ratio and credit risk across Basel eras using EGLS regression on 43 banks. Mandlik and Peshave presented a case-study comparison of NPA management between private and public sector banks. Kumar and Srinivas, in a separate study, applied SVM, Random Forest and Neural Networks for default prediction and RWA optimisation, again finding Random Forest superior (92% accuracy, 90% precision). Dutta and Barman used a Multilayer Perceptron ANN for retail default prediction, and Hussain et al. examined the influence of hedging, diversification and governance on CRM effectiveness. Inekwe studied the macrolevel relationship between banking risk and grain production/GDP, and Bhat and Darzi investigated the effect of rainfall variability on agricultural credit risk, finding it statistically insignificant. Sinha et al. compared Springate, Grover and Zmijewski bankruptcy-prediction models on Indian entertainment companies.

Overall, the literature indicates a clear shift from traditional statistical and survey-based approaches towards machine learning techniques, with ensemble methods such as Random Forest and XGBoost consistently outperforming linear models such as Logistic Regression in predicting credit risk. This motivates the present study, which extends this comparison to a large-scale Indian bank and CIBIL-linked dataset.

III. RESEARCH METHODOLOGY

A. Data Collection and Sample

The study is based on a secondary dataset combining internal bank records and external CIBIL data, comprising 51,336 customer records and 87 features covering demographics, credit score history, loan account details, repayment history, credit enquiries, and bureau credit score. The complete dataset (census method) was used, split into 80% training data

(41,069 records) and 20% testing data (10,267 records). The target variable, Approved_Flag, classifies customers into four risk categories: P1, P2, P3 and P4.

B. Data Preprocessing

The raw data was cleaned by removing missing values and duplicate records, encoding categorical variables into numeric form, and eliminating inconsistent entries. Exploratory Data Analysis (EDA) was then performed using statistical summaries, correlation analysis and visualisations (histograms, box plots and heatmaps) to understand feature distributions and relationships. Relevant predictors were selected using correlation and mutual-information analysis, with Credit_Score, recent enquiry counts (enq_L3m, enq_L6m, enq_L12m), Age_Oldest_TL and personal-loan enquiry ratios emerging as the most informative features.

C. Model Development

Four supervised classification algorithms were implemented using Python, Scikit-learn and XGBoost: Logistic Regression (a linear baseline classifier), Decision Tree (rule-based classification), Random Forest (bagging-based ensemble learning), and Extreme Gradient Boosting – XGBoost (gradient-boosting ensemble with sequential error correction and regularisation). Each model was trained on the 80% training split and evaluated on the held-out 20% test split.

D. Evaluation Metrics

Model performance was assessed using Accuracy,

Precision, Recall, F1-score, Confusion Matrix analysis, and 5fold Cross-Validation, in order to ensure that results were both accurate and generalisable across different subsets of the data.

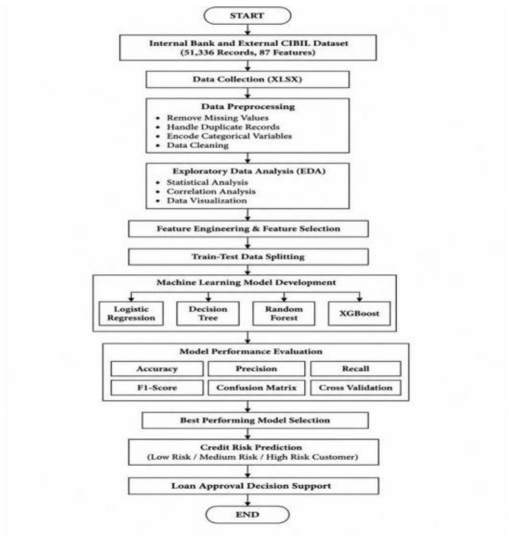

Fig. 1. Proposed methodology: from data collection to credit risk prediction and loan approval decision support.

As illustrated in Fig. 1, the workflow proceeds from data collection (internal bank and external CIBIL dataset) through preprocessing, EDA, feature selection, train-test splitting, model development, performance evaluation, best-model selection, and finally credit risk prediction (Low / Medium / High risk) to support loan approval decisions.

IV. RESULTS AND DISCUSSION

The target variable was found to be imbalanced: P2 accounted for 32,199 records, P3 for 7,452, P4 for 5,882 and P1 for 5,803 records. The customer base was predominantly aged 25–40 years, with credit scores concentrated between 650 and 700 (mean ≈ 680), and male customers (≈ 45,000) far outnumbering female customers (≈ 6,000). Monthly income was strongly right-skewed with a small number of high-income outliers. Graduate-level applicants formed the largest education-level group.

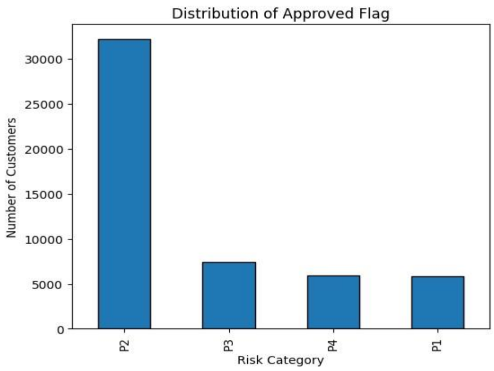

Fig. 2. Distribution of customers across credit risk categories (Approved_Flag).

The correlation heatmap of the top 20 features (Fig. 3) showed mostly weak-to-moderate correlations, with the strongest positive correlations occurring among loan-enquiry and delinquency-related variables, indicating limited multicollinearity and confirming that the selected features contribute largely independent information to the models.

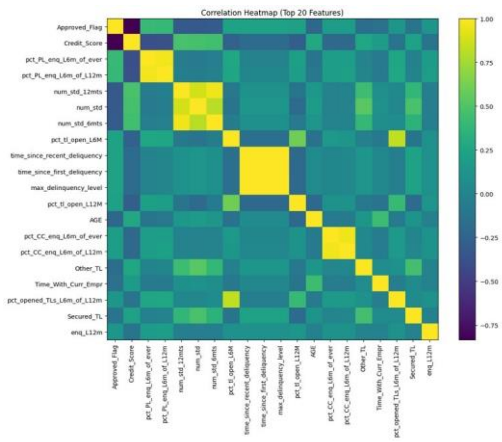

Fig. 3. Correlation heatmap of the top 20 features used in the model.

A. Model Performance

Table I summarises the overall performance of the four models on the test dataset.

Table I. Overall Performance of Machine Learning Models.

Model	Accuracy (%)	F1-Score	Remarks
Logistic Regression	96.65	0.97	Good baseline
Decision Tree	99.15	0.99	Excellent
Random Forest	99.41	0.99	Outstanding
XGBoost	99.48	0.99	Outstanding

XGBoost obtained the highest test-set accuracy (99.48%), followed closely by Random Forest (99.41%), Decision Tree (99.15%) and Logistic Regression (96.65%), as shown in Fig. 4. Precision, Recall and F1-score followed the same pattern (Fig. 5), confirming that the tree-based ensemble methods produced consistently balanced predictions across all four risk categories, while Logistic Regression, though a reliable baseline, was less able to capture the non-linear relationships present in the data.

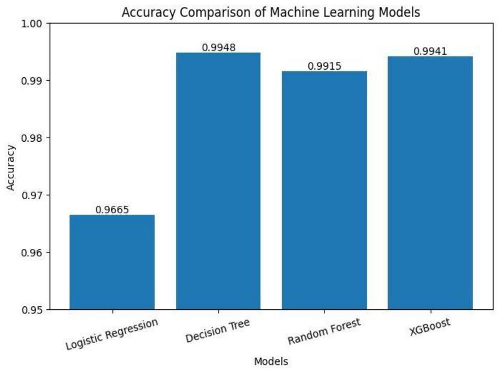

Fig. 4. Accuracy comparison of the four machine learning models.

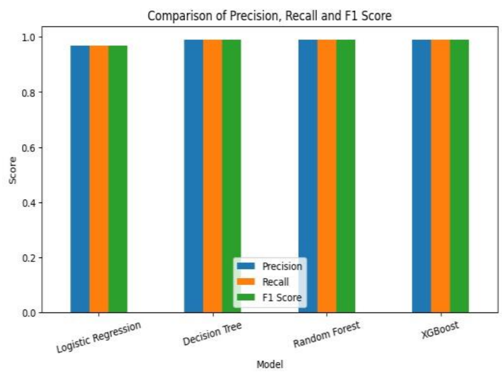

Fig. 5. Comparison of Precision, Recall and F1-Score across models.

B. Cross-Validation Results

Five-fold cross-validation confirmed the robustness of the tree-based models: Decision Tree achieved a mean accuracy of 99.47%, XGBoost 99.43%, Random Forest 99.10%, and Logistic Regression the lowest at 96.40% (Fig. 6). The close agreement between test-set accuracy and cross-validation accuracy for all models indicates good generalisation and a low risk of overfitting, particularly for XGBoost, which combines high accuracy with strong regularisation.

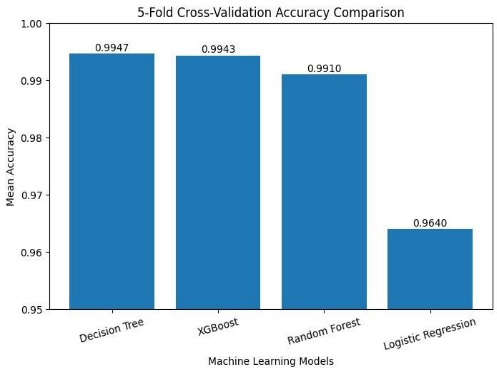

Fig. 6. 5-fold cross-validation accuracy comparison across models.

C. Comparative Evaluation

Table II. Comparative Evaluation of the Four Models.

Criteria	Logistic Regression	Decision Tree	Random Forest	XGBoost
Technique	Linear	Rule-based	Ensemble (bagging)	Ensemble (boosting)
Accuracy	96.65%	99.15%	99.41%	99.48%
Overfitting risk	Low	Moderate	Low	Very low
Interpretability	Excellent	Excellent	Moderate	Moderate

The confusion-matrix analysis showed very few misclassifications for Random Forest and XGBoost, particularly for high-confidence categories, whereas Logistic Regression produced comparatively more errors in the medium-risk (P2/P3) categories. Overall, the results confirm that ensemble learning techniques — particularly gradient boosting — provide superior predictive accuracy, low overfitting risk and strong generalisation for credit risk classification, making them well suited to support automated loan-approval decisions and reduce the growth of NPAs in Indian banks.

V. CONCLUSION

This study developed and compared four machine learning models — Logistic Regression, Decision Tree, Random Forest and XGBoost — for credit risk assessment using a large-scale Indian bank and CIBIL-linked dataset of 51,336 records and 87 features. After data cleaning, encoding and feature selection, all four models achieved high predictive accuracy, with XGBoost (99.48%) and Random Forest (99.41%) outperforming Decision Tree (99.15%) and Logistic Regression (96.65%). Five-fold cross-validation confirmed the robustness of the tree-based ensemble methods. The findings demonstrate that machine learning can substantially improve the speed, objectivity and accuracy of credit risk assessment for Indian Bank, helping identify high-risk borrowers before loan sanction and thereby reducing loan defaults and the growth of Non-Performing Assets. Future work may extend this model with deep learning techniques, real-time data pipelines, and explainability methods such as SHAP to further support transparent lending decisions.

ACKNOWLEDGMENT

The author gratefully acknowledges the guidance of Dr. R N Kulkarni, Professor and Head, Department of Computer Science Engineering, and the support of the Department of Management Studies, Ballari Institute of Technology and Management, Ballari, in carrying out this work.

REFERENCES

1. V. Prabha, R. Kumar, V. Kandpal, and S. Sirsali, “AI-Driven Risk Management to Mitigate Non-Performing Assets in Indian Banks,” 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), 2025.
2. T. K. Kumar and D. Srinivas, “Machine Learning-Driven Evaluation of Credit Risk Models and Provisioning Practices in Indian Banks,” Proc. 5th International Conference on Communication, Computing and Electronics Systems (ICCCES2026), 2026.
3. A. Arora and M. Kumar, “Credit Risk Management Index Score for Indian Banking Sector: An In-Depth Analysis,” The IUP Journal of Bank Management, vol. XIII, no. 1, 2014.
4. L. Koju, R. Koju, and S. Wang, “Does Banking Management Affect Credit Risk? Evidence from the Indian Banking System,” International Journal of Financial Studies, vol. 6, no. 3, 2018.

5. A. Arora, "The Impact of Branch Network on Credit Risk Management Practices in Indian Banking Sector," The IUP Journal of Financial Risk Management, vol. XI, no. 4, 2014.
6. A. Kumar and R. Gupta, "Credit Risk Measurement of Selected Indian Companies," The IUP Journal of Accounting Research & Audit Practices, vol. 22, no. 1, 2023.
7. N. M. Das and J. Deb, "Regulatory Capital and Its Impact on Credit Risk: The Case of Indian Commercial Banks," The IUP Journal of Bank Management, vol. XVI, no. 4, 2017.
8. D. Mandlik and A. D. Peshave, "Credit Risk Management in Indian Banking—A Case Study," Library Progress International, vol. 44, no. 3, pp. 13786–13793, 2024.
9. T. K. Kumar and D. Srinivas, "Evaluation of Credit Risk Management Practices in the Indian Banking Sector Using Machine Learning and Comparative Analysis of Risk Models and Asset Allocation," Proc. 9th International Conference on Inventive Computation Technologies (ICICT-2026), pp. 1953–1960, 2026.
10. A. Dutta and A. N. Barman, "Use of Soft Computing Techniques in Credit Risk Management of an Indian Bank: Application of Artificial Neural Network," 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), pp. 562–567, 2022.
11. A. Hussain, R. Kousar, M. A. Khan, A. B S, K. Ziya, and A. K. S. Ahamed, "Impact of Risk Management Strategies Towards
Credit Risk by Indian Private Banking Companies," 2025 International Conference on Transformative Computing Technologies (ICTCT), pp. 391–396, 2025.
12. J. N. Inekwe, "Bank Credit Risk, Grain Production and the Indian Economy," Applied Economics, vol. 52, no. 47, pp. 5188– 5202, 2020.
13. S. A. Bhat and M. A. Darzi, "Effect of Weather Changes on the Credit Risk in Agricultural Microfinance: An Indian Perspective," Productivity, vol. 59, no. 4, pp. 413–423, 2019.
14. D. Sinha, S. S. Iyer, Y. Rane, and A. Kumar, "Comparative Credit Risk Analysis of Selected Indian Entertainment Companies," The IUP Journal of Accounting Research & Audit Practices, vol. 24, no. 3, pp. 331–354, 2025.