

Predictive Analytics for Credit Risk Assessment in Indian Banks: An Empirical Study of Factors Associated with Loan Defaults

Pratik G. Khandekar*, Dr. Ujwala Narkhede**

*(Department of Management Studies, Wainganga College of Engineering and Management Nagpur, Maharashtra, India

Email: khandekarpratik234@gmail.com)

** (Associate Professor & Research Guide, Wainganga College of Engineering and Management Nagpur, Maharashtra, India.

Email : vaidehi1210@gmail.com)

Abstract:

Credit risk remains the single largest source of losses for the Indian banking system, and the persistence of non-performing assets (NPAs) has renewed interest in data-driven approaches to loan-default prediction. This study develops and evaluates a predictive-analytics framework for credit-risk assessment using a dataset of 1,200 retail and small-business loan accounts drawn from Indian commercial banks. We examine the borrower, loan, and macro-level factors most strongly associated with default and benchmark four classification models: logistic regression, decision tree, random forest, and gradient boosting (XGBoost). Features including credit (CIBIL) score, debt-to-income ratio, loan-to-value ratio, employment stability, and prior delinquency were engineered and evaluated. The gradient-boosting model achieved the strongest discrimination (AUC = 0.91, accuracy = 87.4%), outperforming logistic regression (AUC = 0.82). Debt-to-income ratio, CIBIL score, and prior delinquency emerged as the most influential predictors. The findings suggest that ensemble machine-learning models can materially improve early-warning systems for Indian banks while remaining interpretable through feature-importance analysis. The study offers practical implications for credit-scoring policy, provisioning, and regulatory compliance under the RBI framework.

Keywords — Credit risk, Loan default, Predictive analytics, Machine learning, Indian banks, non-performing assets, Credit scoring

I. INTRODUCTION

The Indian banking sector has undergone rapid credit expansion over the past two decades, driven by retail lending, micro, small and medium enterprise (MSME) financing, and the deepening of financial inclusion. This expansion, however, has been accompanied by a persistent accumulation of non-performing assets (NPAs). Gross NPAs in the Indian banking system peaked at over 11% in the late 2010s before moderating, yet the stock of stressed assets continues to constrain credit growth and erode bank profitability. Accurate assessment of credit risk at the point of loan origination therefore remains a central concern for lenders, regulators, and policymakers.

Traditional credit-appraisal methods in India have relied heavily on judgmental assessment, collateral valuation, and relatively simple scorecards. While bureaus such as CIBIL (TransUnion CIBIL), Experian, and CRIF High Mark now provide standardized credit scores, the underlying scoring logic is often linear and may not capture the complex, non-linear interactions among borrower attributes that drive default. The growing availability of granular loan-level data, combined with advances in machine learning, creates an opportunity to build more accurate and timely early-warning systems.

Predictive analytics - the use of statistical and machine-learning models to forecast future outcomes from historical data - has demonstrated

strong performance in credit scoring across global markets. Yet empirical evidence calibrated to the Indian context, with its distinctive borrower profiles, regulatory environment, and data constraints, remains limited. This study addresses that gap by (i) identifying the borrower-, loan-, and macro-level factors most strongly associated with loan default in Indian banks, and (ii) comparing the predictive performance of conventional and ensemble machine-learning models.

The specific objectives of this study are: (1) to identify and quantify the factors associated with loan defaults; (2) to develop predictive models for default classification; (3) to benchmark model performance using standard evaluation metrics; and (4) to derive practical and policy implications for credit-risk management under the Reserve Bank of India (RBI) regulatory framework. The remainder of the paper is organized as follows. Section II describes the data and methods. Section III explains the preparation of tables. Section IV presents the results and discussion, and Section V concludes.

II. MATERIALS AND METHODS

A. Data Source and Sample

The study uses an anonymized dataset of 1,200 loan accounts compiled from retail and small-business lending portfolios of Indian commercial banks for the period 2019-2023. Each record represents a sanctioned loan and includes borrower demographics, financial attributes, loan characteristics, and the final repayment status. The binary target variable, Default, takes value 1 if the account was classified as a non-performing asset (overdue beyond 90 days per RBI norms) and 0 otherwise. The overall default rate in the sample is 18.3%, reflecting a deliberately stress-weighted portfolio.

B. Variables and Feature Engineering

Predictor variables span three categories. Borrower level: age, CIBIL score, employment stability (months in current employment), and annual income. Loan-level: loan amount, loan-to-value (LTV) ratio, debt-to-income (DTI) ratio, interest rate, and tenure. Behavioural: number of prior delinquencies and credit-utilization ratio. Macro-level: a regional economic-activity index

was merged as a control. Continuous variables were standardized; categorical variables were one-hot encoded. Missing values (below 4% for all fields) were imputed using median (continuous) and mode (categorical) substitution.

C. Models

Four classification models were trained and compared: (1) logistic regression, serving as the interpretable linear benchmark; (2) a single decision tree (CART); (3) a random-forest ensemble; and (4) a gradient-boosting machine (XGBoost). Logistic regression provides directional coefficients and odds ratios, while the tree-based ensembles capture non-linear interactions and yield feature-importance rankings.

D. Model Training and Validation

The data were partitioned into a 70% training set and a 30% hold-out test set using stratified sampling to preserve the default rate. Class imbalance was addressed on the training set using the Synthetic Minority Over-sampling Technique (SMOTE). Hyperparameters were tuned via five-fold cross-validation and grid search. Model performance was assessed on the untouched test set using accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). All modelling was performed in Python using the scikit-learn and XGBoost libraries.

III. PREPARATION OF TABLES

Tables in this paper follow the IEEE two-column convention. Each table is numbered with uppercase Roman numerals in the order of its first citation in the text, and the caption is placed above the table in 8-point small capitals. Table cell text is set in 9-point Times New Roman, with column headers in bold. Wide tables that cannot fit within a single column are permitted to span both columns at the top or bottom of a page. Explanatory notes, where required, appear immediately below the table in 8-point italic type. Three tables are presented: Table I summarizes the descriptive statistics of the sample; Table II reports the logistic regression estimates of the factors associated with default; and Table III compares the predictive performance of the four models.

Numeric precision is standardized across tables - coefficients and ratios are reported to two or three decimal places, and statistical significance is denoted by conventional asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). This standardization ensures that the tables are self-contained and interpretable without reference to the surrounding text.

IV. RESULTS AND DISCUSSION

A. Descriptive Statistics

Table I presents the descriptive statistics of the principal variables for the full sample and separately for defaulters and non-defaulters. Defaulters exhibit markedly lower CIBIL scores, higher debt-to-income ratios, and a greater incidence of prior delinquency than non-defaulters, providing initial univariate evidence of their predictive relevance.

TABLE I
DESCRIPTIVE STATISTICS BY REPAYMENT STATUS

Variable	All	Non-default	Default
CIBIL score	726	748	651
DTI ratio (%)	38.2	34.1	56.7
LTV ratio (%)	71.4	69.0	82.1
Annual income (Rs. lakh)	7.9	8.4	5.6
Prior delinquencies	0.6	0.3	1.9
Loan tenure (months)	58	57	61

Values are group means. Rs. lakh = INR 100,000. sample ($n = 1,200$).

B. Factors Associated with Default

Table II reports the logistic-regression estimates. Debt-to-income ratio, CIBIL score, prior delinquency, and loan-to-value ratio are all statistically significant at conventional levels. A higher CIBIL score is associated with lower odds of default (odds ratio < 1), whereas higher DTI and LTV ratios and a history of delinquency sharply increase default odds. Employment stability is protective, consistent with the intuition that stable income supports repayment capacity.

TABLE II
LOGISTIC REGRESSION ESTIMATES OF LOAN DEFAULT

Predictor	Coef. (b)	Odds ratio	Sig.
CIBIL score	-0.021	0.979	***
DTI ratio	0.058	1.060	***
LTV ratio	0.034	1.035	**
Prior delinquency	0.712	2.038	***
Employment stability	-0.014	0.986	*
Annual income	-0.045	0.956	**
Constant	2.914	-	**

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

C. Predictive Performance

Table III compares the four models on the hold-out test set. The gradient-boosting model (XGBoost) delivered the best overall performance, with an AUC of 0.91 and accuracy of 87.4%, followed closely by the random forest. Logistic regression, while the most interpretable, trailed the ensembles on every metric, reflecting its inability to capture non-linear interactions among predictors. The superior recall of the ensemble models is particularly valuable for lenders, since correctly identifying likely defaulters (true positives) is more consequential than occasionally flagging a sound borrower.

TABLE III
COMPARISON OF MODEL PERFORMANCE ON TEST SET

Model	Acc.	Prec.	Recall	F1	AUC
Logistic regression	0.804	0.72	0.68	0.70	0.82
Decision tree	0.821	0.75	0.73	0.74	0.84
Random forest	0.861	0.83	0.80	0.81	0.89
Gradient boosting	0.874	0.86	0.83	0.84	0.91

Metrics computed on the 30% hold-out test set.

D. Discussion

The results reinforce three findings relevant to Indian banks. First, a small set of variables - DTI ratio, CIBIL score, prior delinquency, and LTV ratio - accounts for most of the predictive signal, implying that even data-constrained lenders can build effective early-warning scores. Second, ensemble machine-learning models offer a meaningful accuracy gain over linear scorecards without sacrificing interpretability, since feature-importance analysis preserves transparency for internal audit and regulatory scrutiny. Third, the strong role of behavioural variables (prior delinquency, utilization) argues for richer integration of bureau and transaction data into origination workflows.

These findings carry practical implications. Banks can embed gradient-boosting scores into loan-origination systems to triage applications, calibrate risk-based pricing, and inform provisioning under the expected-credit-loss

paradigm. From a regulatory standpoint, interpretable ensemble models align with RBI expectations for model governance and explainability. Limitations include reliance on a single cross-sectional dataset and the exclusion of unstructured data (e.g., alternative digital-footprint signals), which future research could incorporate.

CONCLUSIONS

This study developed and evaluated a predictive-analytics framework for credit-risk assessment in Indian banks. Using a dataset of 1,200 loan accounts, it identified debt-to-income ratio, CIBIL score, prior delinquency, and loan-to-value ratio as the dominant factors associated with loan default. Among the four models tested, gradient boosting achieved the strongest discrimination (AUC = 0.91, accuracy = 87.4%), clearly outperforming the logistic-regression benchmark. The evidence indicates that ensemble machine-learning models can strengthen early-warning systems and credit-scoring practices while remaining interpretable through feature-importance analysis.

For practitioners, the results support the integration of machine-learning scores into loan origination, risk-based pricing, and provisioning. For regulators, interpretable ensemble models offer a path toward more accurate yet governable credit models. Future work should extend the analysis to longitudinal data, incorporate alternative and transactional data sources, and test model stability across economic cycles and borrower segments.

ACKNOWLEDGMENT

The author gratefully acknowledges the guidance of the faculty of the Department of Management Studies and the support of the participating institutions in facilitating access to anonymized loan data. The author also thanks colleagues and reviewers for their constructive feedback on earlier drafts of this manuscript.

REFERENCES

[1] Reserve Bank of India, "Report on Trend and Progress of Banking in India," RBI, Mumbai, 2023.

- [2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [3] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785-794.
- [4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.
- [5] D. West, "Neural network credit scoring models," *Computers & Operations Research*, vol. 27, no. 11-12, pp. 1131-1152, 2000.
- [6] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring," *European J. Operational Research*, vol. 247, no. 1, pp. 124-136, 2015.
- [7] B. Baesens et al., "Benchmarking classification techniques for credit scoring," *J. Operational Research Society*, vol. 54, no. 6, pp. 627-635, 2003.
- [8] TransUnion CIBIL, "Understanding Your Credit Score," TransUnion CIBIL Ltd., Mumbai, 2022.
- [9] A. Khandani, A. Kim, and A. Lo, "Consumer credit-risk models via machine-learning algorithms," *J. Banking & Finance*, vol. 34, no. 11, pp. 2767-2787, 2010.
- [10] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.