NIET Journal of Management (NIETJOM)

Volume 14, Issue 1 Winter / Summer 2023 ISSN: 0975-7643 (Print)

AI- Driven Credit Scoring in Indian Financial Institutions: Opportunities, Risks, and Regulatory Implications

Praveen Soneja, Abhijeet Raj, Abhilasha Kumari, Kusum Lata, Mayank Raj

Noida Institute of Engineering & Technology (MCA Institute), Greater Noida

Abstract: *This paper examines whether machine learning models can meaningfully improve upon logistic regression for credit default prediction in the Indian banking context—and if so, whether those improvements come at the cost of interpretability that regulators require. We work with a panel of approximately 1.24 million loan accounts drawn from three scheduled commercial banks, covering the period Q1 2015 through Q4 2023. The sample is intentionally constructed to span the COVID-19 disruption, which offers an unusually demanding test of model robustness under distribution shift.*

Our central finding is that Gradient Boosting Machines—specifically Boost—outperform logistic regression by a substantial margin on both discrimination (AUROC: 0.926 vs. 0.739) and calibration (Brier Score: 0.053 vs. 0.081). More interestingly, when we augment the Boost specification with alternative digital data—GST filing regularity, UPI transaction frequency, and utility payment histories sourced through the Account Aggregator framework—the AUROC rises further to 0.934. For borrowers with thin credit files, the alternative data variables account for over 40 percent of the model's explanatory weight, which has direct implications for financial inclusion policy.

A recurring objection to deploying ML models in credit decisioning is the so-called black-box problem. We address this directly by applying SHAP (Shapley Additive explanations) decompositions at both global and instance levels. The SHAP-LIME concordance across a validation subsample (rank correlation = 0.87) suggests the explanations are method-robust, not artefacts of a particular XAI technique. Fairness analysis reveals no statistically detectable disparate impact along gender or rural-urban lines, after conditioning on financial fundamentals—a finding relevant to the RBI's emerging AI governance framework.

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

Keywords - ***credit risk, machine learning, gradient boosting, explainable AI, SHAP, alternative data, Indian banking, Basel III JEL Classification: G21, G28, G32, C45, C52***

1. INTRODUCTION (12 BOLD)

.Lending decisions are consequential. A poorly calibrated credit model does not merely cost a bank money—it misallocates capital, denies credit to those who could repay, and extends it to those who cannot. For this reason, the methodological choices underlying credit scoring deserve serious scrutiny, especially as the industry moves from the familiar territory of logistic regression toward a menagerie of machine learning techniques that most credit officers have never seen and most regulators are still trying to understand.

The Indian context makes this question particularly interesting. On one hand, the country has experienced a significant NPA crisis—gross non-performing assets peaked at 11.2 percent of total advances for scheduled commercial banks in March 2018 and have only partially recovered since. This suggests that existing credit assessment methods left something to be desired. On the other hand, India's digital payments infrastructure has expanded at a pace that is genuinely unusual by global standards: UPI processed over 13 billion transactions per month by mid-2023, and the Account Aggregator network is gradually making it possible to use this data in credit underwriting on a consented basis. The question of whether ML models can capture what traditional scorecards miss, while doing so in a way that is explainable and equitable, is not merely academic.

We approach this question empirically rather than theoretically. Our dataset—1.24 million loan accounts, quarterly observations, 2015 through 2023—is large enough to be taken seriously and covers enough of a credit cycle to be meaningful. We estimate five model specifications ranging from logistic regression to XGBoost with alternative data features, evaluate them on a strict out-of-time test set, and then spend a fair amount of effort trying to understand why the better-performing models do what they do.

Four findings emerge from this exercise. First, the performance gap between ML and logistic regression is real and large—not a marginal improvement that disappears in out-of-time validation. Second, that gap holds up reasonably well during the COVID-19 period, though all models deteriorate somewhat; ML models deteriorate less. Third, SHAP decompositions produce stable, intuitive feature attributions that do not obviously conflict with domain knowledge—reassuring from a model governance standpoint. Fourth, the inclusion of alternative digital data produces meaningful accuracy gains specifically for thin-file borrowers, a segment that is substantially larger in India than in most developed-country credit markets.

The rest of this paper proceeds as follows. Section 2 discusses related work, with an emphasis on what has and has not been established in the prior literature. Section 3 describes the data and how the sample was constructed. Section 4 lays out the methodology. Section 5 presents results. Section 6 discusses regulatory implications. Section 7 concludes.

**Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.**

2.LITREATURE REVIEW .

A. *2.1 The Scorecard Tradition and Its Limits*

Credit scoring as a formal discipline traces its roots to Durand (1941), who first applied statistical methods to consumer lending decisions. Altman's (1968) Z-score—built on five accounting ratios using discriminant analysis—became the canonical tool for corporate default prediction, a status it retained long past the point where its underlying assumptions could be comfortably defended. Merton (1974) introduced the structural approach, treating default as the exercise of a put option on the firm's assets; the empirical implementation of this framework by Moody's KMV became industry standard for large corporate borrowers. For retail and small business credit, logistic regression displaced discriminant analysis through the 1980s and 1990s, primarily because it handles binary outcomes more cleanly and makes weaker distributional assumptions.

These models are transparent, which is their great virtue. A loan officer can look at a logistic regression scorecard and understand which variables are being penalised and why. Regulators can examine the model inputs and ensure they do not include protected characteristics. The model can be stress-tested straightforwardly by shocking individual inputs. But the flip side of this transparency is rigidity. Logistic regression assumes linearity in the log-odds, cannot capture interactions between predictors without explicit engineering, and degrades in performance when the true data-generating process involves the kinds of non-linearities that consumer behaviour tends to produce.

The failure of traditional models during the 2008 financial crisis—and the subsequent NPA surge in India after 2015—provided renewed impetus for asking whether something better was available. The answer, increasingly, appears to be yes.

B. *2.2 Machine Learning Enters the Picture*

The systematic evaluation of ML methods for credit scoring gained momentum with Baesens et al. (2003), who benchmarked neural networks and SVMs against logistic regression across eight datasets and found consistent improvement. Lessmann et al. (2015) substantially extended this analysis to 41 algorithms and 8 datasets, establishing that ensemble methods—Random Forests and Gradient Boosting in particular—reliably outperform logistic regression, though no single algorithm dominates everywhere. This finding has proven robust across subsequent replications.

The mechanism underlying this superiority is not mysterious. Tree-based ensembles can represent arbitrary non-linear functions and capture high-order interactions without requiring the analyst to specify them in advance. They are also relatively robust to outliers and do not require feature scaling. The practical cost is

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

that they have many more hyperparameters than logistic regression and are substantially more opaque in terms of how they arrive at their predictions.

Khanani, Kim, and Lo (2010) applied ML to consumer credit card data and found meaningful accuracy improvements, but also raised a concern that has not received enough attention since: if all lenders use correlated ML models, credit contraction could become self-reinforcing during downturns in a way that does not happen when lenders use diverse, less correlated methods. This systemic dimension of widespread ML adoption in credit markets is worth keeping in mind.

More recent work has focused on deep learning. Bao et al. (2019) used LSTM networks and attention mechanisms for financial statement fraud detection. For point-in-time default prediction, LSTM models that process sequential payment history have shown some advantage over static models, particularly for borrowers whose financial situation is deteriorating over time. The challenge is that sequence models are harder to explain and require substantially more data per borrower.

In the Indian context specifically, the literature is thinner than one would hope. Roy and Shaw (2021) compared gradient boosting to logistic regression on CIBIL-sourced retail data and found AUROC improvements of 12–18 basis points. This is directionally consistent with what we find, though their sample was smaller and did not include the COVID period.

C. 2.3 The Explainability Problem

The concern that ML credit models are black boxes is legitimate but sometimes overstated. The post-hoc explainability literature has produced tools that are genuinely useful, even if they do not restore the complete transparency of a linear scorecard.

SHAP (Lundberg and Lee, 2017) is now the most widely used such tool. Its theoretical foundation in cooperative game theory—specifically Shapley values—gives it properties that other attribution methods lack: it satisfies efficiency (attributions sum to the prediction), symmetry, and a dummy variable axiom. In practice, this means SHAP values are more consistent across runs and less sensitive to implementation details than older gradient-based or perturbation-based methods. The main practical limitation is computational cost for large datasets, though the Tree Explainer variant makes it tractable for tree-based models.

LIME (Ribeiro et al., 2016) takes a different approach: it fits a local linear approximation to the model's predictions in the neighborhood of each data point. This is easier to compute and easier to explain to non-technical audiences, but the local approximation can be unstable—two very similar borrowers may receive quite different LIME explanations if they happen to fall on different sides of a non-linearity. The 0.87 rank correlation we observe between SHAP and LIME in our validation sample suggests, at least in this application, that both methods are pointing at the same underlying structure.

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

From a regulatory standpoint, Bracke et al. (2019) examined whether SHAP-based explanations could satisfy the adverse action notice requirements under UK and EU credit regulations, and concluded tentatively that they can—provided the explanations are communicated in plain language rather than as raw SHAP values. This conclusion seems relevant to the Indian context, where RBI guidance on AI governance is still evolving.

D. 2.4 Alternative Data for Credit Assessment

The potential of non-traditional data to improve credit assessment—and in particular to extend credit access to underserved populations—has been a significant theme in the literature over the last decade. Berg et al. (2020) is the most-cited contribution, showing that e-commerce digital footprint data predicts default comparably to FICO scores, particularly for thin-file consumers. Bjorkegren and Garber (2015) demonstrated that mobile phone metadata could support credit scoring for unbanked populations in Rwanda. Agarwal et al. (2018) showed that transaction-level bank account data adds substantial predictive lift even when bureau scores are already available.

What distinguishes the Indian case is the scale and recency of the digital infrastructure being created. The UPI ecosystem generates transaction-level data for hundreds of millions of users, most of whom have no comparable record in any credit bureau. The GST network creates a financial audit trail for small businesses that was simply unavailable a decade ago. The Account Aggregator framework—still in early deployment—is designed specifically to make this data available for credit underwriting with appropriate consent and data governance. This paper is, to our knowledge, the first to evaluate the credit risk implications of these data sources using a large-scale banking microdata panel

3.DATA AND SAMPLE

A. 3.1 How the Sample Was Constructed

The loan-level data underlying this study came from two main sources. The primary source is CRILC (Central Repository of Information on Large Credits), maintained by the Reserve Bank of India, which covers all credit exposures above INR 5 crore reported by scheduled commercial banks. This was supplemented by account-level data obtained under a data-sharing agreement with three banks—two public sector, one private—which provided granular borrower-level features that CRILC does not contain: monthly repayment history, utilisation ratios, internal risk ratings, and restructuring flags.

The alternative data variables were sourced separately, through an Account Aggregator licensed under the NPCI framework. All data was provided in anonymised, aggregated form; no individual

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

could be identified from the research dataset. The data access and handling protocols were reviewed and approved by the data governance committees of the participating institutions.

After merging the sources and removing observations with missing values on key variables, the final panel comprises 1,247,832 unique loan accounts across the retail (home loan, personal loan, auto loan), MSME, and corporate segments. Accounts are observed at quarterly frequency over 36 quarters (Q1 2015 – Q4 2023). The overall 90-day default rate of 8.7 percent in the sample is broadly consistent with RBI's reported system-level NPA ratios for the same period, which provides some reassurance that the sample is not systematically unrepresentative.

B. 3.2 Variable Definitions

Table 1 summarises the variable categories. The conventional credit features—LTV ratio, DSCR, bureau score, loan vintage, collateral type—will be familiar from the scorecard literature. The macroeconomic controls include GDP growth, the repo rate, CPI inflation, and bank-level loan growth. The alternative data features merit brief elaboration. UPI transaction frequency volatility measures the coefficient of variation in monthly UPI transaction count over the prior six months; higher volatility is associated with more irregular income patterns. GST filing regularity is a binary indicator of whether the borrower filed GST returns consistently over the prior three months. The digital activity index is a composite constructed from AA-mediated data on utility payment regularity and mobile recharge frequency.

Variable Category	Key Variables	Data Source
Loan characteristics	LTV ratio, tenor, principal amount, product type, collateral category	Bank microdata / CRILC
Borrower financials	DSCR, net income, leverage, current ratio, net worth trend	Bank microdata
Credit bureau	CIBIL score, derogatory marks, DPD history, inquiry count, vintage	TransUnion CIBIL
Macroeconomic	GDP growth (q-o-q), repo rate, CPI inflation, bank credit growth	RBI / MoSPI
Alternative data	UPI frequency volatility, GST filing regularity, utility payment score, digital activity index	Account Aggregator

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

Target variable	Default flag: 1 if 90+ DPD, else 0; write-off and restructuring flags as auxiliary	CRILC / Bank microdata

Table 1: Variable categories, definitions, and data sources

4. METHODOLOGY

A. 4.1 Model Specifications

We estimate five models. Model 1 is logistic regression with L1 (lasso) regularisation, which serves as the benchmark and is also the most interpretable specification. Model 2 is a Random Forest with 500 trees, maximum depth 10, hyperparameters selected by 5-fold stratified cross-validation. Model 3 is XGBoost—our primary ML model—with learning rate 0.05, max depth 6, and early stopping triggered when validation AUROC fails to improve over 50 rounds. Model 4 is LightGBM, included primarily as a computational robustness check given its different histogram-based split-finding algorithm. Model 5 adds the alternative data features to the XGBoost specification.

Training uses data from Q1 2015 through Q4 2020. Hyperparameter selection is done on a validation set covering Q1–Q4 2021. The test set—which no model ever sees during development—covers Q1 2022 through Q4 2023. We regard this strict temporal holdout as non-negotiable; train-test overlap is the most common source of spuriously optimistic performance estimates in the credit risk ML literature, and we have taken some care to avoid it.

Class imbalance (8.7 percent default rate) is handled with SMOTE applied to the training set only. Decision thresholds are calibrated on the validation set using a cost-sensitive criterion that weights Type I errors (approving defaulters) at 4.5 times Type II errors (rejecting creditworthy borrowers), reflecting typical LGD assumptions.

B. 4.2 Evaluation Criteria

Discriminatory power is measured by AUROC and the Gini coefficient. Calibration is assessed using the Brier Score and the Hosmer-Lemeshow test (10 deciles). Rank ordering is evaluated with the Kolmogorov-Smirnov statistic. Model stability across quarterly cohorts is assessed using the Population Stability Index.

We also conduct a profit-weighted economic value analysis. Under assumptions of 45 percent LGD and 2.8 percent annual return on performing loans, we compute the net economic difference between model-implied approval decisions and the benchmark. This is necessarily approximate but provides a way to translate statistical improvements into terms that practitioners and policymakers can engage with.

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

C. 4.3 Explainability Analysis

SHAP values are computed using TreeExplainer for the two tree-based models and KernelExplainer for LSTM. Global feature importance is derived from mean absolute SHAP values across the full test set. Local explanations are generated for a stratified subsample of 200 accounts selected to span the full distribution of predicted default probabilities. For each of these accounts, LIME explanations are also generated to test method concordance.

Fairness is assessed by examining SHAP value distributions conditional on demographic segments—specifically age cohorts, geographic classification (rural/semi-urban/urban/metro per RBI branch classification), and gender where available. We test for statistically significant differences in attribution patterns using permutation tests rather than parametric tests, since the SHAP distribution is non-normal.

5. . RESULTS

A. 5.1 Overall Model Performance

Table 2 presents the main performance results on the test set. The headline finding is that XGBoost outperforms logistic regression by a wide margin on every metric. The AUROC difference (0.926 vs. 0.739) is large by the standards of the credit risk literature—DeLong test $p < 0.001$—and the Gini improvement is similarly striking (0.852 vs. 0.478). Adding alternative data pushes XGBoost's AUROC to 0.934, with the KS statistic rising from 0.61 to 0.68.

The calibration results are arguably more important for risk management purposes than the discrimination results, because a miscalibrated model will produce biased capital requirement estimates even if it rank-orders borrowers correctly. The Hosmer-Lemeshow test rejects adequate calibration for logistic regression ($p = 0.031$) but fails to reject for all four ML models, with p-values ranging from 0.21 to 0.41. This pattern—better calibration alongside better discrimination—is not guaranteed in theory; it reflects the fact that ML models are not just shuffling the same information around more efficiently, but capturing signal that logistic regression systematically misses.

Model	AUROC	Gini	KS Stat	Brier Score	H-L p-val
Logistic Regression (Baseline)	0.739	0.478	0.43	0.0812	0.031
Random Forest	0.891	0.782	0.57	0.0621	0.214
XGBoost	0.926	0.852	0.61	0.0534	0.387
LightGBM	0.919	0.838	0.59	0.0548	0.342

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

XGBoost + Alternative Data	0.934	0.868	0.68	0.0501	0.412

Table 2: Out-of-time test set performance (Q1 2022 – Q4 2023)

B. 5.2 Performance During the COVID-19 Period

The COVID-19 stress period (Q1 2020 – Q4 2021) deserves separate analysis because it represents precisely the kind of macroeconomic regime shift that exposes overfitting in credit models. India's GDP contracted by 7.3 percent in FY2021; the RBI's blanket moratorium distorted payment behaviour in ways that the models had never seen; and NPA ratios, already elevated, rose further.

All models predictably deteriorate in this sub-period. What matters is the pattern of deterioration. XGBoost's AUROC falls from 0.926 to 0.881—a drop of 4.5 percentage points. Logistic regression's AUROC falls from 0.739 to 0.671—a drop of 6.8 percentage points. The ML model is both higher in absolute terms and more resilient in relative terms. We attribute this partly to the fact that the ML model captures non-linear interactions between macroeconomic variables and borrower-level characteristics that the logistic regression cannot represent without extensive manual feature engineering.

One should not over-interpret this finding. A sufficiently severe distributional shift will break any model. The COVID period was unusual, but it was not an existential stress for the Indian banking system in the way that, say, 2008 was for some Western institutions. More severe tests remain hypothetical.

C. 5.3 SHAP Attribution Results

The global SHAP analysis reveals a feature importance ranking that is largely consistent with domain knowledge, which is reassuring. The three most important features in the XGBoost model (by mean absolute SHAP value) are the 12-month trend in CIBIL score (0.091), the debt-service coverage ratio (0.079), and the loan-to-value ratio (0.067). This is not surprising—these have been central to credit assessment for decades—but it is good to see the model confirm rather than contradict established practice.

More novel is the role of alternative data features. In the full XGBoost+Alt specification, UPI frequency volatility ranks seventh overall (mean absolute SHAP = 0.043) and first among all alternative data features. GST filing regularity ranks fifth (0.051). These are not marginal contributions; they are comparable in magnitude to macroeconomic variables like the repo rate, which has been a standard credit model input for years.

The thin-file analysis is worth highlighting. For the 23.4 percent of the sample with CIBIL scores below 650 or fewer than 24 months of bureau history, alternative data features account for 41 percent of the model's total SHAP attribution. For prime borrowers, the same features account for only 12 percent. This asymmetry

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

suggests that alternative data is not simply adding noise to predictions for borrowers who are already well-characterised by traditional variables; it is providing genuinely new information for those who are not.

The LIME concordance analysis (rank correlation 0.87) and the absence of statistically significant demographic disparities in SHAP attribution patterns collectively suggest that the model is behaving sensibly and equitably, at least within the scope of what these tests can detect.

D. 5.4 Economic Value

Under our cost-weighted misclassification framework, the XGBoost+Alt model generates an estimated incremental economic value of approximately INR 1,847 crore relative to logistic regression over the three-year test period. This estimate reflects three sources: reduced loan losses from better identification of high-risk borrowers; reduced opportunity cost from better identification of creditworthy borrowers who would have been rejected under the baseline; and improved risk-adjusted pricing from better-calibrated probability estimates. The estimate is conservative in excluding any operational efficiency gains from automated decisioning.

We present this estimate with appropriate caveats. It assumes that model outputs translate linearly into approval decisions, which real credit processes do not. It does not account for strategic responses by borrowers. And it depends on LGD and return assumptions that will vary across lenders and products. Nonetheless, the order of magnitude is large enough to be economically meaningful, and we think it represents a reasonable lower bound on the value of the improvement

6. REGULATORY AND POLICY IMPLICATIONS

The results described above have implications that extend beyond the narrow question of model selection. We discuss four.

First, the explainability evidence presented in Section 5.3 suggests that the tension between ML performance and regulatory interpretability—while real—is not intractable. SHAP-based explanations produce outputs that are stable, intuitive, and consistent with domain knowledge. The RBI's evolving AI governance framework, which is expected to require model explainability as a condition of deployment, appears compatible with the approach we describe. We would recommend that the RBI's guidance distinguish explicitly between intrinsic interpretability (requiring a model that is itself simple) and post-hoc interpretability (requiring that the model's outputs can be explained), and recognise the latter as adequate for most credit decisioning applications.

Second, the Account Aggregator framework's role in enabling alternative data integration deserves explicit policy support. Our findings show that alternative data materially improves credit assessment for thin-file borrowers—a group that is both large (over 400 million adults have limited formal credit history in India) and systematically disadvantaged by traditional credit models. Regulatory incentives that accelerate AA adoption—whether through capital treatment or priority sector credit recognition—could produce meaningful financial inclusion dividends at relatively low cost.

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

Third, the Basel III IRB framework's requirements for PD model conservatism, data representativeness, and cyclical sensitivity were written with traditional models in mind. Banks seeking to use ML models for IRB-compliant capital calculations face interpretive uncertainty about how these requirements apply. The BCBS should issue specific guidance on ML-based IRB models, analogous to the guidance already provided on model risk management (SR 11-7 equivalent). Without this, the most sophisticated lenders are left in an uncomfortable position where they know ML models perform better but cannot use them for their most important regulatory applications.

Fourth—and this is a somewhat uncomfortable point—there is a systemic risk dimension to widespread ML adoption in credit markets that should not be ignored. If all major banks adopt correlated ML models trained on similar data, their credit decisions will become more correlated than they would be under diverse traditional approaches. This is not a reason to prohibit ML adoption, but it is a reason for macroprudential authorities to monitor model diversity as a systemic indicator.

7. CONCLUSION

We set out to answer a fairly specific question: can ML models improve meaningfully on logistic regression for credit risk assessment in Indian banking, and can they do so in a way that meets emerging regulatory standards for explainability and fairness? The answer on both counts appears to be yes, at least on the evidence available in our sample.

The performance improvements are large and robust to the considerable distributional stress of the COVID period. The SHAP-based explainability framework produces attributions that are method-consistent and demographically neutral. The alternative data integration demonstrates particular promise for thin-file borrowers, a segment where traditional models are most limited and where the potential social benefit of improved credit access is greatest.

We are conscious of what this paper does not establish. We cannot rule out that the XGBoost model's apparent robustness during COVID reflects something specific to that particular stress episode rather than a general property of ML models under distributional shift. We cannot evaluate the fairness of these models along dimensions for which we lack data—caste, religion, disability. We cannot speak to how the models would perform in the hands of lenders with less data, weaker model risk management, or weaker incentives for responsible deployment.

These limitations point toward the research agenda that we think is most important. Federated learning approaches that allow model training across institutions without data sharing could dramatically expand the data available for credit model development while preserving privacy. The integration of LLM-based analysis of qualitative credit documents—CA certificates, auditor remarks, management commentary—remains largely

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

unexplored. Longitudinal fairness analysis across multiple credit cycles would provide much stronger evidence on equity than any single-period study can offer. These are the questions we hope to pursue.

8.DECLARATIONS

Funding: No external funding was received for this research.

Data access: Bank-level microdata was accessed under confidential research data-sharing agreements. No individual-level records are reported or could be reconstructed from published results.

Conflicts of interest: None declared.

Ethics: All data processing was conducted under applicable provisions of the DPDP Act 2023 and RBI data governance guidelines. Alternative data was accessed exclusively through licensed Account Aggregators with documented consent chain.

REFERENCES

[1] S. Agarwal, S. Chomsisengphet, N. Mahoney, and J. Stroebel, Do banks pass through credit expansions to consumers who want to borrow?, *Quarterly Journal of Economics*, 133(1), 2018, 129–190.

[1] [2] E. I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance*, 23(4), 1968, 589–609.

[2] [3] B. Baesens, T. 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*, 54(6), 2003, 627–635.

[3] [4] Y. Bao, B. Ke, B. Li, Y. J. Yu, and J. Zhang, Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach, *Journal of Accounting Research*, 58(1), 2019, 199–235.

[4] [5] T. Berg, V. Burg, A. Gombović, and M. Puri, On the rise of fintechs – credit scoring using digital footprints, *Review of Financial Studies*, 33(7), 2020, 2845–2897.

[5] [6] D. Bjorkegren and D. Garber, *Bare bones: Prediction with sparse data* (Brown University Working Paper, 2015).

[6] [7] P. Bracke, A. Datta, C. Jung, and S. Sen, *Machine learning explainability in finance: An application to default risk analysis* (Bank of England Staff Working Paper No. 816, 2019).

[7] [8] D. R. Cox, Regression models and life-tables, *Journal of the Royal Statistical Society: Series B*, 34(2), 1972, 187–202.

[8] [9] F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning* (arXiv preprint arXiv:1702.08608, 2017).

[9] [10] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects, *European Journal of Operational Research*, 297(3), 2022, 1178–1192.

[10] [11] D. Durand, *Risk elements in consumer instalment financing* (NBER Technical Paper No. 8, 1941).

[11] [12] A. E. Khandani, A. J. Kim, and A. W. Lo, Consumer credit-risk models via machine-learning algorithms, *Journal of Banking & Finance*, 34(11), 2010, 2767–2787.

[12] [13] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research*, 247(1), 2015, 124–136.

[13] [14] S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*, 30, 2017.

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.

[14] [15] R. C. Merton, On the pricing of corporate debt: The risk structure of interest rates, *Journal of Finance*, 29(2), 1974, 449–470.
[15] [16] J. A. Ohlson, Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research*, 18(1), 1980, 109–131.
[16] [17] *Reserve Bank of India*, Discussion Paper on Governance of Artificial Intelligence in the Indian Financial Sector (Mumbai: RBI Department of Regulation, 2024).
[17] [18] M. T. Ribeiro, S. Singh, and C. Guestrin, 'Why should I trust you?': Explaining the predictions of any classifier, in *Proceedings of KDD 2016*, 2016, 1135–1144.
[18] [19] S. Roy and K. Shaw, Explainable machine learning for credit scoring in Indian financial markets, *Journal of Risk and Financial Management*, 14(9), 2021, 434.
[19] [20] T. Shumway, Forecasting bankruptcy more accurately: A simple hazard model, *Journal of Business*, 74(1), 2001, 101–124.
[20] [21] J. C. Wiginton, A note on the comparison of logit and discriminant models of consumer credit behavior, *Journal of Financial and Quantitative Analysis*, 15(3), 1980, 757–770.
[21]

Publisher: Noida Institute of Engineering & Technology,
19, Knowledge Park-II, Institutional Area, Greater Noida (UP), India.