Bincang Sains dan Teknologi (BST)
DOI: 10.56741/bst.v5i03.2758

E-ISSN 2961-8746
P-ISSN 2961-8932



AI-Driven Credit Scoring and Loan Default Prediction

A Comparative Machine Learning Study of Credit Risk Assessment and Lending Decisions

[1]T. Venkatesan*, [1]A. Stephen, [1]S. Lakshana, [2]Ozal Saeed, [2]Asghar Khan
Corresponding Author: * proftvv@gmail.com
[1] Knowledge Institute of Technology, Salem, Tamil Nadu, India
[2] Abasyn University, Peshawar, Pakistan

Article submitted: 5 October 2026
Revised: 13 November 2026
Accepted: 15 November 2026

Abstract

Credit default prediction is central to consumer lending because it informs credit approval, portfolio monitoring, pricing, and provisioning. This quantitative comparative study examined the predictive performance of traditional credit-scoring approaches and artificial intelligence (AI)-based models using secondary financial lending data covering the stated study period of 2015–2025. The research considered LendingClub, Home Credit Default Risk, UCI German Credit, and macroeconomic information from Federal Reserve Economic Data. Logistic regression and a FICO-based scorecard were compared with random forest, extreme gradient boosting (XGBoost), and a long short-term memory (LSTM) network. The analytical procedure included data cleaning, imputation, normalization, feature selection, class-imbalance treatment, an 80:20 training–testing split, and fivefold cross-validation. The performance comparison used accuracy, AUC-ROC, F1-score, RMSE, and R^2. XGBoost yielded the strongest reported results, with accuracy of 91.2%, AUC-ROC of 0.96, and F1-score of 0.89, followed by LSTM and random forest. The SHAP feature-ranking analysis identified repayment history, debt-to-income ratio, employment length, credit utilization, and recent inquiries as leading predictors. The findings demonstrate stronger descriptive discrimination by AI-based approaches in the evaluated comparison, while indicating that model calibration, cross-dataset harmonization, fairness, and out-of-time validation remain important considerations. The study contributes comparative evidence to the application of machine learning in retail credit assessment and discusses implications for responsible lending and financial inclusion.

Keywords: *Banking Risk Management, Credit Scoring, Default Prediction, Machine Learning, Random Forest, XGBoost.*

Introduction

Credit risk assessment is fundamental to financial intermediation because banks and other lenders make lending decisions before observing whether individual borrowers will repay their obligations. The consequences of an inaccurate assessment extend beyond the classification of an application: underestimated default probabilities may increase credit losses and capital pressures, whereas overly restrictive assessments can exclude potentially creditworthy customers. Credit scoring consequently links statistical prediction with financial decision-making, pricing, provisioning, loan monitoring, and supervisory expectations [1], [2]. The objective is not merely to maximize the number of correct classifications, but to develop a sufficiently reliable account of repayment risk for decisions under uncertainty.

Traditional systems have commonly applied credit-bureau information, repayment histories, financial ratios, scorecards, and multivariate statistical methods. Altman established an influential early example of multivariate distress classification, and subsequent consumer-credit literature formalized statistical approaches to scoring and applicant selection [1], [2]. Logistic regression remains a principal benchmark because it provides a tractable mapping between financial characteristics and estimated event probabilities. However, a logit model with only additive linear predictors may fail to represent threshold effects and interactions unless these features are explicitly modeled. Lending records often contain heterogeneous borrowers, uneven data quality, and nonlinear patterns that motivate consideration of flexible algorithms [3], [4].

Machine learning has expanded the range of classification strategies. Random forests aggregate multiple decision trees, while gradient boosting fits ensembles sequentially to reduce prediction loss [5], [6]. XGBoost adds regularization and computational efficiency to gradient-boosted trees and is particularly relevant to high-dimensional tabular records [7]. LSTM networks can summarize temporal dependencies in sequences through gated recurrent units, potentially supporting applications when repeated borrower behavior is observed over time [8]. The use of these methods should be assessed empirically rather than presupposed to be superior: the strongest algorithm may vary with sample size, the performance horizon, class imbalance, feature definitions, and the data available at the time of the lending decision [3], [4].

Credit scoring also raises interpretability and distributional questions. Explainability methods such as SHAP can identify features associated with a prediction, while fairness research recognizes that historical lending data may reproduce unequal access to credit [9]–[12]. Alternative information sources could enlarge credit coverage among borrowers without extensive bureau records, although their benefits depend on validity, consent, and the absence of inappropriate proxy discrimination [13], [14]. Evaluating predictive performance without considering these practical constraints would provide an incomplete basis for implementation.

The present study comparatively evaluates logistic regression, a FICO-based scorecard, random forest, XGBoost, and LSTM for loan-default prediction. It also examines the relative importance of explanatory variables and the implications of predictive scoring for loan origination, credit-risk controls, and access to financial services. Three research questions guide the analysis: (i) how do AI-based methods compare with the traditional models in classification performance; (ii) which borrower characteristics contribute most prominently to the machine-learning predictions; and (iii) what do these findings imply for responsible credit decision-making? The contribution is an empirical comparison using the available observed results, supplemented by methodological interpretation relevant to banking practice.

Materials and Methods

A. Research design and data sources

A quantitative comparative research design was adopted to analyze binary loan-default outcomes. The stated period of analysis was 2015–2025. The financial information was obtained from publicly

available lending datasets and macroeconomic resources. LendingClub supplied the broader lending-record context, with approximately 2.26 million records reported for 2007–2020 and more than 150 candidate fields. The Home Credit Default Risk repository comprised 307,511 application records, and the UCI German Credit dataset comprised 1,000 observations. Federal Reserve Economic Data supplied macroeconomic contextual variables. These data sources reflect distinct lending environments and therefore provide heterogeneous financial and demographic characteristics rather than one inherently interchangeable sampling frame. The study considered repayment and credit histories, debt-to-income ratios, income, credit utilization, loan amounts, employment characteristics, inquiries, and behavioral variables. The response of primary interest was default status, understood as the classification outcome used in the respective loan datasets. An important analytical distinction is that source datasets do not necessarily adopt identical default horizons, operational definitions, or reporting conventions. Accordingly, interpretation of a combined credit-risk score requires careful consideration of label compatibility, date alignment, and eligible predictor fields. The descriptive statistics and predictive results are presented using the values obtained in this study; the reported data do not specify a final record-level pooled denominator by source or the exact cross-source linkage and harmonization rules. Standardized feature-selection and regularization concepts are discussed in general statistical-learning references [15], [16].

B. Data preprocessing and feature selection

Prior to modeling, the data-processing procedures comprised duplicate removal, screening for inconsistent observations, missing-value treatment through median imputation or interpolation, and standardization using StandardScaler. A correlation-based examination and recursive feature elimination were applied to identify relevant variables and limit redundant inputs. Because default and nondefault observations may be unequally represented, the Synthetic Minority Oversampling Technique (SMOTE) was included in the preprocessing strategy. SMOTE creates synthetic minority observations from nearby examples and is widely used in imbalanced classification [17]. Preprocessing is consequential for interpretation. Income, debt ratios, loan sizes, and utilization measures can differ substantially in their numerical scales, and some algorithms are more sensitive to these differences than tree-based models. Missingness may also encode information about documentation quality or the structure of a credit application. The meaning of imputation consequently depends on why values are missing, whether they were available when the lender made the decision, and whether transformations are estimated in the training partition. In a replicable workflow, feature selection, imputation, scaling, and oversampling should be confined to the training folds to prevent information leakage into validation and test predictions [18], [19]. The study focused on predictors that can be interpreted in lending terms: repayment history reflects demonstrated performance, debt-to-income ratio relates debt obligations to earning capacity, credit utilization measures reliance on available revolving facilities, and recent credit inquiries may reflect additional borrowing demand. Employment length offers a proxy for income continuity but can also create concerns about differences in employment structure. Model explanations must be interpreted as statistical associations rather than proof that changing one feature independently changes the probability of default.

C. Comparative model specifications

Logistic regression and a FICO-based scorecard served as conventional comparators. Logistic regression models logit[P(Y=1|X)] as a linear combination of the included predictor terms, and the fitted probability is classified using a specified decision cutoff. Scorecard methods use structured borrower information and risk-related point assignments; they are attractive in lending because rules and reason codes can be communicated to credit analysts [2], [20]. Linear discriminant analysis was also discussed among the candidate statistical methods, but the available performance table includes numerical results for logistic regression and the FICO scorecard only; therefore, no result is attributed to an unreported discriminant-analysis model. Random forest constructs a collection of trees using resampling and randomized predictor selection, with class predictions or estimated probabilities aggregated across the ensemble [5]. XGBoost sequentially builds trees to optimize a differentiable loss function with regularization [6], [7]. These methods can capture complex dependencies and interactions without requiring a single globally linear predictor structure. LSTM employs recurrent gates to preserve or update information across sequences [8], [21]. In credit scoring, meaningful LSTM use requires properly ordered borrower observations or repayment histories rather than merely arranging unrelated application fields into an artificial sequence. The study describes an LSTM architecture with three stacked layers, dropout of 0.3, and a sigmoid output for binary prediction.

D. Training, tuning, and evaluation

The dataset was divided into an 80% model-development partition and a 20% test partition. Fivefold cross-validation was used during model development. Grid search was specified for traditional models, while Bayesian optimization was used for the ensemble and deep-learning methods. Python 3.11 provided the machine-learning workflow, including scikit-learn, XGBoost, TensorFlow/Keras, SHAP, LIME, Pandas, and NumPy. R packages and IBM SPSS/Excel were also identified as validation and analytical resources. The reported comparison is based on the resulting accuracy, AUC-ROC, F1-score, RMSE, and R^2 metrics. Accuracy denotes the fraction of predictions classified correctly at the chosen threshold. Sensitivity or recall is the fraction of actual defaults correctly detected, while precision is the fraction of predicted defaults that actually default. F1-score summarizes precision and recall through their harmonic mean. ROC-AUC evaluates ranking over thresholds and is useful for comparing the discriminatory ability of scoring systems [22]. In situations where the default class is relatively uncommon, precision–recall curves provide additional information beyond ROC curves [23], [24]. RMSE summarizes squared differences between observed outcomes and predictions; its interpretation in a binary setting depends on whether scores represent probabilities or classification labels. R^2 is reported as an additional metric in the analysis, although its definition must be distinguished from the conventional variance-explanation measure used in continuous-outcome regression. The methodological framework emphasizes consistent model comparison. Performance indices are comparable only if they arise from the same evaluation observations, target definitions, and appropriate information sets. An out-of-time holdout is particularly relevant to banking because future borrowers and macroeconomic conditions may differ from those represented during development. Cross-validation can improve tuning but does not eliminate risks from time leakage, dependence among

repeated loans, or population drift [18], [19], [25]. The numerical comparisons below are therefore interpreted as the observed model ranking, while uncertainty around these estimates cannot be quantified without the underlying loan-level predictions.

E. Model explainability, fairness, and research ethics

Explainability was examined using Shapley Additive Explanations (SHAP), which assigns feature contributions relative to a chosen reference distribution [9]. Local Interpretable Model-agnostic Explanations (LIME) was also included in the stated analytical tools [10]. The feature ranking was based on mean absolute SHAP values, which describe the magnitude of model contributions across observations. These quantities should not be interpreted as causal effects, and a global ranking can hide variation between individual borrowers. Fairness was considered within the framework of privacy, discrimination risk, transparency, and post-processing adjustments; however, numerical results for specific demographic groups and fairness metrics are not part of the available results tables. The study analyzed existing secondary financial data rather than recruiting participants directly. Nevertheless, financial data can contain potentially identifying or sensitive attributes, requiring careful attention to source permissions, secure handling, and regulatory conditions. Credit decisions are high stakes: even accurate predictions may disadvantage particular borrower groups when historical selection processes are reproduced. Model governance therefore should evaluate both predictive errors and the distribution of approvals, rejections, and adverse consequences across relevant groups [11], [12], [26].

Results

A. Descriptive characteristics of the credit data

Table 1 summarizes the principal borrower and loan variables. The average debt-to-income ratio was 18.40% (SD 9.20), with values ranging from 0.00% to 45.80%. Mean annual income was 62.50 in the reported thousand-rupee scale (SD 38.10), while average loan amount was 14.80 (SD 8.60) in the corresponding thousand-rupee notation. Mean credit utilization reached 55.20%, with an SD of 28.40 percentage points. The default-rate variable averaged 18.70%, with a range of 8.10% to 34.20%. These values describe variation in financial capacity, borrowing exposure, and outcome prevalence within the study summary.

Table 1. Descriptive statistics of the analyzed credit variables.

Variable	Mean	Standard Deviation	Minimum	Maximum
Debt-to-Income Ratio (%)	18.40	9.20	0.00	45.80
Annual Income (₹ 000s)	62.50	38.10	12.00	250.00
Credit Utilisation (%)	55.20	28.40	0.00	99.80
Loan Amount (₹ 000s)	14.80	8.60	1.00	40.00
Default Rate (%)	18.70	5.30	8.10	34.20

The dispersion of income, utilization, and loan amounts indicates that the credit records encompass a broad range of applicant circumstances. A large utilization range, for example, permits a model to represent materially different levels of use of revolving credit facilities. However, aggregate descriptive statistics do not reveal which individual features are jointly distributed among defaulters. Consequently, the predictive models were assessed through multivariable combinations of borrower characteristics

rather than conclusions drawn from any single mean. The default-rate measure is presented as originally estimated; its unit of aggregation is not further specified.

B. Comparative model performance

Table 2 presents the five model results. XGBoost achieved the highest accuracy (0.912), AUC-ROC (0.96), F1-score (0.89), and R^2 (0.88), alongside the lowest RMSE (0.11). LSTM followed with accuracy of 0.895, AUC-ROC of 0.94, F1-score of 0.87, and RMSE of 0.13. Random forest achieved accuracy of 0.887 and AUC-ROC of 0.93, with F1-score of 0.86 and RMSE of 0.14. Both AI-based ensembles and LSTM exceeded the values reported for logistic regression and the FICO-based scorecard.

Table 2. Comparative prediction performance of traditional and AI-based credit-scoring models.

Model	Accuracy	AUC-ROC	F1-Score	RMSE	R^2 Score
Logistic Regression	0.784	0.81	0.75	0.21	0.65
FICO Scorecard	0.749	0.76	0.70	0.24	0.60
Random Forest	0.887	0.93	0.86	0.14	0.82
XGBoost	0.912	0.96	0.89	0.11	0.88
LSTM	0.895	0.94	0.87	0.13	0.84

The accuracy difference between XGBoost and logistic regression was 0.128, equivalent to 12.8 percentage points. The corresponding difference between XGBoost and the FICO scorecard was 0.163, or 16.3 percentage points. On AUC-ROC, XGBoost exceeded logistic regression by 0.15 and the FICO scorecard by 0.20. These comparisons indicate a consistent ordering across the tabulated indices, with XGBoost leading the AI-based group. The table does not provide confidence intervals, standard errors, paired test statistics, or a common confusion matrix. Therefore, these differences represent observed performance contrasts rather than independently demonstrated statistical significance.

The reported F1-scores also favor AI-based models, suggesting a potentially useful balance of default-class precision and recall under the applied classification thresholds. Their managerial implications depend on the actual classification thresholds and cost distribution of false decisions. The inclusion of RMSE and R^2 supplies another comparison but does not replace probability calibration or the evaluation of default-specific losses. Strong discrimination is an important result for credit ranking; accurate probability estimation is a related but separate operational requirement.

C. Traditional and AI-based scoring comparison

Table 3 summarizes the comparison of conventional and AI-based approaches across performance and functional characteristics. Traditional models were classified as exhibiting moderate accuracy of approximately 74%–78%, while the AI-based models achieved approximately 88%–91%. The corresponding AUC-ROC ranges were 0.76–0.81 and 0.93–0.96. Additional dimensions included error rates, risk identification, data handling, and adaptability. The reported categories reflect the methodological comparison accompanying the numerical analysis, with “high” and “low” representing qualitative rather than separately estimated statistical outcomes.

The distinction between methods has practical implications. Traditional models can remain attractive when transparency, small training samples, or implementation costs are priorities. AI-based models provide greater flexibility for heterogeneous input features and nonlinear dependencies. Nevertheless, the ability to adapt to new financial conditions depends on retraining and validation procedures; it is not achieved automatically by selecting an algorithm.

Table 3. Comparison of traditional and AI-based credit-scoring approaches.

Criteria	Traditional Models	AI-Based Models
Prediction Accuracy	Moderate (74–78%)	High (88–91%)
AUC-ROC Score	0.76 – 0.81	0.93 – 0.96
Error Rate	High	Low
Risk Identification	Limited	Comprehensive
Data Handling	Structured only	Structured + Alternative
Adaptability	Static	Dynamic / Adaptive

The findings thus favor AI-based performance under the tested comparison without excluding the continued relevance of conventional models as benchmarks or simpler operational options.

D. Explainable AI and predictor importance

The SHAP-based analysis identified five prominent risk-related predictors (Table 4). Repayment history had the highest mean absolute contribution, at 0.34, followed by debt-to-income ratio (0.28), employment length (0.19), credit utilization rate (0.17), and number of recent inquiries (0.12). The stated directional interpretation associates better repayment histories and longer employment with lower predicted default risk, and higher debt ratios, utilization, and inquiry counts with higher predicted risk.

Table 4. SHAP feature-importance analysis of loan-default prediction.

Rank	Feature	Mean |SHAP| Value	Direction
1	Repayment History	0.34	Negative (higher = lower risk)
2	Debt-to-Income Ratio	0.28	Positive (higher = higher risk)
3	Employment Length	0.19	Negative (longer = lower risk)
4	Credit Utilisation Rate	0.17	Positive (higher = higher risk)
5	Number of Recent Inquiries	0.12	Positive (more = higher risk)

The ranking is substantively consistent with the purpose of retail credit assessment: both historical repayment and contemporaneous debt burden provide information about repayment capacity and behavior. The relative importance of employment length suggests that continuity of income-related circumstances may contribute to discrimination, although its interpretation can differ by employment type and country. A mean absolute attribution describes importance within a fitted model, not the magnitude of a change in true default probability caused by altering that feature. In particular, repayment-history information must predate the score date to be a valid origination predictor. These considerations affect how the model should be explained to credit officers and applicants.

E. Implications for lending and risk management

The comparative results indicate several applications for banking decisions. First, improved discrimination can help identify high-risk applications for additional review or appropriately constrained approval. Second, borrower-level scoring can support differentiated pricing when predicted probabilities are calibrated and coupled with loss-given-default and exposure information. Third, combining conventional variables with appropriately governed alternative data may help evaluate applicants with limited formal histories. Fourth, portfolio aggregation can assist risk monitoring and provisioning. These implications follow from the predictive task but should be distinguished from measured reductions in realized loan losses, which were not separately evaluated in the study.

E-ISSN 2961-8746
P-ISSN 2961-8932

Discussion

A. Interpretation of machine-learning performance

The empirical comparison shows that XGBoost had the strongest overall performance on the reported metrics. This result accords with the ability of gradient-boosted trees to represent irregular nonlinear relationships among financial predictors and their interactions [6], [7]. For example, a debt-to-income ratio may be associated with different default risk depending on existing repayment history or utilization. An additive conventional logit specification would require these interactions to be explicitly modeled, while tree methods may identify them through recursive partitioning. The advantage found here should be interpreted within the evaluated modeling exercise rather than assumed to hold identically across countries, time periods, and lending products.

Random forest also showed strong predictive results, reflecting the utility of aggregating unstable single trees into a more robust ensemble [5]. LSTM performed slightly better than random forest on the tabulated accuracy and AUC-ROC measures. Its architecture may be relevant when sequential repayment or transaction information is available [8], [21]. However, the comparative table alone does not explain the performance mechanism, and additional sequence-level ablation analysis would be needed to attribute the LSTM result specifically to its representation of temporal dependencies. These patterns underscore the importance of comparing modeling approaches using consistent input information rather than treating architectural sophistication as a guarantee of improvement.

Previous credit-scoring benchmarks have also found that no single classifier should be designated universally optimal and that evaluation design affects algorithm ordering [3], [4]. The present results contribute another example in which ensemble methods perform competitively. They also show the importance of maintaining a conventional comparator, because accuracy and AUC differences become informative when tied to a defined baseline. For practical use, lenders should examine how robust these differences remain under changed economic conditions and among borrower subgroups.

B. Predictive discrimination, calibration, and economic consequences

A high AUC-ROC indicates that a model effectively distinguishes between defaulting and nondefaulting cases across decision thresholds. Nevertheless, discrimination alone does not show whether predicted probabilities correspond to observed event frequencies. Two models may rank applicants similarly while assigning very different numerical probabilities. Calibration assessment, including reliability plots and Brier scores, is important when credit prices, capital allocations, or expected-loss calculations depend on probability levels [27], [28]. The reported RMSE values provide complementary error information but cannot substitute for a documented calibration analysis.

A lending institution also needs to choose a threshold at which scores translate into approvals, rejections, manual review, or differentiated pricing. This decision depends on expected revenue, cost of funding, operational expenses, loss given default, and regulatory constraints. A false approval can create a financial loss, whereas a false rejection may forgo profitable lending and deny an otherwise creditworthy applicant access to finance. Because these costs differ, an accuracy-maximizing threshold need not be the economically or socially appropriate decision threshold [29]. A cost-sensitive evaluation

would be a valuable next step for connecting the present classification gains with actual business performance.

Credit scoring also operates in a portfolio rather than exclusively at the level of individual applications. Systematic shocks can induce correlated losses even when application-level ranking is strong. Models trained on historical borrowers may perform differently after unemployment changes, interest-rate shocks, or transformations of underwriting policies. This makes repeated out-of-time testing, drift monitoring, and comparison against established risk-management policies essential [25], [30]. The result that XGBoost performed best in the evaluated sample is thus a starting point for deployment assessment, not a substitute for lifecycle management. The wider challenges of time-dependent financial modeling and data-rich financial prediction have also been discussed in the literature [31], [32].

C. Feature interpretation and financial behavior

The feature-ranking analysis highlights familiar dimensions of creditworthiness. Repayment history is prominent because past performance may provide information about future willingness and capacity to repay. Debt-to-income ratio measures current financial commitments relative to income. Credit utilization may reveal financial pressure or strong reliance on revolving facilities. Recent inquiries may be associated with changing credit demand, although the direction and magnitude of this association can depend on product type and borrower circumstances. Employment length can reflect stability but may also carry structural differences unrelated to individual repayment intention.

SHAP analysis contributes to model transparency by identifying features that materially affect the fitted score [9]. It can guide checks on unexpected relationships and support internal model review. However, a SHAP value does not establish a causal mechanism or guarantee an acceptable adverse-action explanation. Correlated features may share attribution in different ways, and a variable important globally need not dominate the explanation for a particular applicant [10], [33]. Credit analysts should consequently examine local predictions, model stability, and reason-code consistency alongside the aggregate importance ranking.

D. Fairness and financial inclusion

AI-assisted scoring may expand the set of observable borrower signals and offer opportunities for applicants with limited conventional credit histories. Studies of digital footprints and alternative data illustrate how such information can complement standard records, especially when traditional scoring coverage is incomplete [13], [14]. However, improved prediction is not equivalent to expanded access. Inclusion depends on how the score changes actual approval decisions, loan conditions, and subsequent repayment outcomes. It is also necessary to examine whether nontraditional variables reflect uneven access to technology, location-based disadvantages, or historically unequal lending practices.

The study incorporated fairness and ethical considerations within its methodological scope, but no numerical subgroup comparison is displayed. Thus, the results support a discussion of the need for fairness assessment rather than a numerical conclusion that disparity has been reduced. An evaluation could examine approval rates, false-positive and false-negative rates, and calibration across legally appropriate borrower groups [11], [12], [34]. Different fairness criteria can lead to different

classifications of a model as equitable, and mitigation strategies may involve trade-offs among predictive utility, access, and error distribution. These questions are operationally important in the deployment of credit-scoring technology.

The regulatory context requires particular attention. Model documentation should make clear which data can be collected, how long they may be retained, what explanations are available to applicants, and how adverse decisions can be contested. Banking risk management frameworks emphasize the need for risk controls, independent validation, and responsibility for model use [35], [36]. The European AI Act and the NIST AI Risk Management Framework provide additional reference points for risk-based governance and documentation, although specific legal obligations depend on jurisdiction [37], [38].

E. Methodological strengths and limitations

The study contributes by comparing several important traditional and AI-based credit classifiers with common summary metrics and by incorporating a feature-importance perspective. Its use of an 80:20 development–testing structure, cross-validation, and systematic hyperparameter tuning follows the broad logic of predictive model assessment. The range of data sources also highlights the practical heterogeneity of consumer credit information. Reporting multiple indices rather than accuracy alone is beneficial because classification involves several kinds of prediction error.

Several limitations qualify the interpretation. The available study information does not specify the exact harmonization of default outcomes and borrower records across the named datasets or a final analytic sample size. There is also no detailed account of whether the data split was chronological, whether repeated loans from the same borrower were kept within a single partition, or whether oversampling and imputation were performed exclusively within each training fold. Leakage, sampling dependence, and changes in the distribution of default risk can affect generalization [18], [19]. These issues do not negate the reported performance results but reduce confidence in extending them beyond the analyzed comparison.

A second limitation is the absence of uncertainty intervals or paired model-comparison tests accompanying Table 2. The numerical performance differences can be described precisely, but statistical significance cannot be inferred solely from differences in accuracy or AUC. Likewise, fairness improvement and changes in actual nonperforming-loan ratios would require subgroup analyses and realized postdecision outcomes. Future studies should report class distributions, test denominators, confusion matrices, calibration, temporal holdout results, subgroup metrics, and model specifications. A third limitation is that the qualitative attributes of the comparison in Table 3 are not independent outcome measurements; they should be read in conjunction with the numerical results.

F. Future research directions

Future research should test the generality of the model ordering on clearly defined separate lending cohorts and independent out-of-time samples. Distinct benchmark datasets can first be analyzed separately with a comparable pipeline; any pooled modeling should then document compatible target definitions and feature mappings. This approach would help distinguish a robust algorithmic result from a gain driven by one dataset. A second line of inquiry should investigate calibration and cost-sensitive

approval policies rather than discriminatory performance alone. An economically grounded evaluation could quantify expected losses and forgone revenue under transparent assumptions.

A third direction is to integrate interpretable decision-support tools with fairness audits, applicant communication, and model governance. Research could test whether credit officers make better calibrated decisions when provided with feature explanations and uncertainty estimates. Longitudinal work would be required to determine whether automated scoring improves portfolio delinquency, financial inclusion, or risk-adjusted performance after implementation. Finally, the role of open-banking and digital behavioral data should be assessed with careful attention to privacy, legal permission, and unequal access. Such work would extend the present findings from predictive classification toward accountable lending systems.

Conclusion

This study compared traditional credit-scoring approaches with machine-learning models for loan-default prediction. The observed performance rankings favored XGBoost, which achieved accuracy of 91.2%, AUC-ROC of 0.96, and F1-score of 0.89. LSTM and random forest also recorded stronger performance than logistic regression and the FICO-based scorecard on the reported metrics. The SHAP analysis identified repayment history, debt-to-income ratio, employment length, credit utilization, and recent inquiries as influential predictors. Together, these findings demonstrate the relevance of flexible nonlinear models to credit-risk classification and the value of explainability for understanding model outputs.

The results have practical significance for banks considering more granular underwriting and portfolio risk assessment. However, their implementation value depends on calibrated probabilities, consistent data definitions, appropriate thresholds, reliable validation, fair treatment of borrowers, and governance throughout the model lifecycle. Future studies with transparent sample construction, model-level predictions, subgroup results, and observed lending outcomes can determine how the predictive advantages translate into reduced credit losses and more inclusive decisions. AI-based credit scoring should therefore be evaluated as an integrated financial decision process, not solely as a competition in classification accuracy.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] E. I. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” The Journal of Finance, vol. 23, no. 4, pp. 589–609, 1968, doi: 10.1111/j.1540-6261.1968.tb00843.x.

[2] D. J. Hand and W. E. Henley, “Statistical classification methods in consumer credit scoring: A review,” Journal of the Royal Statistical Society: Series A, vol. 160, no. 3, pp. 523–541, 1997.

[3] A. E. Khandani, A. J. Kim, and A. W. Lo, “Consumer credit-risk models via machine-learning algorithms,” Journal of Banking & Finance, vol. 34, no. 11, pp. 2767–2787, 2010, doi: 10.1016/j.jbankfin.2010.06.001.

[4] 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, vol. 247, no. 1, pp. 124–136, 2015, doi: 10.1016/j.ejor.2015.05.030.

[5] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[6] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
[7] 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, doi: 10.1145/2939672.2939785.
[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[11] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," European Journal of Operational Research, vol. 297, no. 3, pp. 1083–1094, 2022, doi: 10.1016/j.ejor.2021.06.023.
[12] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA, USA: MIT Press, 2023.
[13] T. Berg, V. Burg, A. Gombović, and M. Puri, "On the rise of FinTechs: Credit scoring using digital footprints," The Review of Financial Studies, vol. 33, no. 7, pp. 2845–2897, 2020, doi: 10.1093/rfs/hhz099.
[14] V. B. Djeundje, J. Crook, R. Calabrese, and M. Hamid, "Enhancing credit scoring with alternative data," Expert Systems with Applications, vol. 163, Art. no. 113766, 2021, doi: 10.1016/j.eswa.2020.113766.
[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed. New York, NY, USA: Springer, 2021.
[16] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
[18] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proc. 14th Int. Joint Conf. Artificial Intelligence, 1995, pp. 1137–1143.
[19] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, Art. no. 91, 2006, doi: 10.1186/1471-2105-7-91.
[20] L. C. Thomas, D. B. Edelman, and J. N. Crook, Credit Scoring and Its Applications. Philadelphia, PA, USA: SIAM, 2002.
[21] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," European Journal of Operational Research, vol. 270, no. 2, pp. 654–669, 2018, doi: 10.1016/j.ejor.2017.11.054.
[22] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, 1997, doi: 10.1016/S0031-3203(96)00142-2.
[23] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. 23rd Int. Conf. Machine Learning, 2006, pp. 233–240, doi: 10.1145/1143844.1143874.
[24] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
[25] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Computing Surveys, vol. 46, no. 4, Art. no. 44, 2014, doi: 10.1145/2523813.
[26] World Bank, The Global Findex Database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID-19. Washington, DC, USA: World Bank, 2022, doi: 10.1596/978-1-4648-1897-4.
[27] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Machine Learning, 2005, pp. 625–632, doi: 10.1145/1102351.1102430.
[28] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.

[29] C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th Int. Joint Conf. Artificial Intelligence, 2001, pp. 973–978.
[30] J. C. Hull, Risk Management and Financial Institutions, 5th ed. Hoboken, NJ, USA: Wiley, 2018.
[31] R. S. Tsay, Analysis of Financial Time Series, 3rd ed. Hoboken, NJ, USA: Wiley, 2010.
[32] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," The Review of Financial Studies, vol. 33, no. 5, pp. 2223–2273, 2020, doi: 10.1093/rfs/hhaa009.
[33] C. Molnar, Interpretable Machine Learning, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[34] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems, vol. 29, 2016.
[35] Basel Committee on Banking Supervision, Principles for the Management of Credit Risk. Basel, Switzerland: Bank for International Settlements, 2000.
[36] Basel Committee on Banking Supervision, International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Basel, Switzerland: Bank for International Settlements, 2006.
[37] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence," Official Journal of the European Union, 2024.
[38] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.

Authors



Dr. T. Venkatesan is an Associate Professor and Head of the Department of Management Studies at Knowledge Institute of Technology (Autonomous), Salem, Tamil Nadu, India. He has over 18 years of experience in teaching, research, and academic administration, specializing in Finance and Accounting. He holds a Ph.D. in Management Studies (Finance) from Bharathiar University and has published more than 56 research articles in national and international journals, including Scopus- and Web of Science-indexed journals. His research interests include Corporate Finance, Financial Management, Working Capital Management, Banking, and Investment Analysis. He has also served as Vice-Principal, guided research scholars, and contributed significantly to management education and academic research. (email: proftvv@gmail.com).



Dr. A. Stephen is the Director of the Knowledge Institute of Technology, Salem, Tamil Nadu, India. His research focuses on human resource management, financial literacy, and consumer behavior. He has published studies on credit card promotion schemes, challenges faced by women entrepreneurs, and financial literacy among young people in Tamil Nadu, contributing to social science and management research. (email: directorkbs@kiot.ac.in).



S. Lakshana is an MBA candidate specializing in Finance and Marketing at Knowledge Institute of Technology, Salem, India. She holds a Bachelor of Commerce degree in Computer Applications. She gained practical experience in SAP-based commercial operations, vendor management, and financial analysis through an internship at JSW Steel (Salem Works), as well as marketing experience at Aavin. Her professional skills include financial analysis, market research, Tally, and Power BI. She is interested in applying her analytical and strategic capabilities to business development, financial management, and organizational performance improvement. (email: slakshana1905@gmail.com).



Dr. Ozal Saeed is a Ph.D. scholar at Abasyn University, Peshawar. His academic interests focus on interdisciplinary research within management, education, and public administration. Committed to academic excellence, he is involved in various scholarly activities and research collaborations aimed at contributing to the advancement of knowledge in his field. (email: ozalsaeed@gmail.com).

E-ISSN 2961-8746
P-ISSN 2961-8932



Dr. Asghar Khan is a PhD scholar at Abasyn University, Peshawar, with a strong academic and professional background in public sector finance and administration. He is currently serving as Assistant Director (Accounts) at the Election Commission of Pakistan, where he is engaged in financial management, regulatory compliance, and institutional governance. His academic interests and professional experience focus on public financial management, accountability, and the effective functioning of democratic institutions. (email: asgharmusazai@gmail.com).