Journal of Machine Learning and Deep Learning (JMLDL)
Volume 3, Issue 1, Date: 30- June – 2026
ISSN: 2978-4336, DOI: 10.64820/AEPJMLDL.31.29.38.62026

Deep Learning vs. Traditional Machine Learning for Consumer Credit Default Prediction: A Comparative Study

Nguyen Trung Hieu, MBA Candidate
Executive Business Centre, University of Greenwich, London, England, United Kingdom · Alliance with FPT University, Ho Chi Minh City, Viet Nam
Email: nh7471m@gre.ac.uk
*Corresponding Author

***Abstract*—Credit default prediction is fundamental to risk management in consumer lending. This study compares logistic regression (LR), a multilayer perceptron (MLP), and XGBoost for credit default prediction using 30,000 synthetic consumer lending records modeled after LendingClub data (34.0% default rate, 22 features). Stratified 5-fold cross-validation reveals that LR achieves the highest AUC-ROC (0.726 ± 0.006), marginally exceeding MLP (0.723 ± 0.006) and XGBoost (0.705 ± 0.007). Bootstrap DeLong tests confirm all pairwise AUC differences are significant ($p < 0.001$), and the Friedman test validates the overall ranking ($\chi^2 = 10.0$, $p = 0.007$). However, XGBoost achieves the highest F1-score (0.548) and recall (0.584), detecting substantially more defaults at the cost of lower precision. LR achieves the best calibration (Brier score = 0.192). These results illustrate a metric-dependent trade-off: LR ranks best by discrimination (AUC) and calibration, while XGBoost ranks best by default detection rate. The findings caution practitioners against equating model complexity with predictive value and argue for metric-driven model selection in regulated lending.**

Keywords—credit scoring; default prediction; logistic regression; gradient boosting; neural network; model comparison; probability calibration

I. INTRODUCTION

Credit scoring models guide lending decisions that affect millions of consumers and support the stability of consumer-credit markets [1]. A scoring model estimates the probability that a borrower will default, and this estimate drives approval, pricing, and provisioning decisions. Even small gains in predictive accuracy turn into large monetary effects at portfolio scale, which is why the choice of modeling technique remains a recurring and important question for lenders.

Consumer lending also operates under tight supervisory oversight. Capital rules require each lender to estimate a probability of default for every exposure, and supervisors expect these estimates to be stable, well documented, and explainable to both regulators and rejected applicants. A model that lifts a headline accuracy figure but cannot be explained, monitored, or reproduced is of limited practical use, so any modeling choice must be judged against operational and regulatory constraints, not discrimination alone.

The regulatory backdrop shapes which models are practical. Under the internal-ratings-based approach of the Basel framework, a lender estimates a probability of default for each exposure and derives capital requirements from it, so the model output must be a well-calibrated probability, not merely a rank order [1]. Supervisory validation then examines the stability of the model, the meaning of its inputs, and the reasons given for each decline, which favors models whose behavior can be traced from input to output [2]. These requirements make calibration and interpretability first-class objectives, so the study reports a calibration metric and a feature-importance analysis, not discrimination alone.

Logistic regression has been the industry standard since the 1980s because it is interpretable, easy to monitor, and aligned with regulatory needs such as the Basel accords and adverse-action explainability rules [1], [2]. Over the past decade, a growing body of work has reported that machine learning and deep learning methods improve discrimination on credit data [3]-[5]. The evidence is mixed. Gains are often small, inconsistent across data sets, and rarely tied to a clear statement of which evaluation metric the gain appears under [6]-[8]. This leaves practitioners without a clear answer to a basic question: does added model complexity actually pay off for consumer credit default prediction, and under which evaluation criterion?

Resolving this question matters for three reasons. First, complex models carry real costs, including reduced interpretability, harder regulatory validation, and higher computational and maintenance overhead, so added complexity must earn a clear, measurable benefit. Second, credit data are highly structured and the relationship between features and default is often close to monotonic, a setting in which simple linear models are hard to beat [1]. Third, comparisons that report a single headline metric can mislead, because discrimination, calibration, and detection rate need not move together. A controlled comparison that holds the data fixed and evaluates several complementary metrics is therefore needed.

Received: 3-3- 2026
Revised: 277-6-2026
Published: 30-6-2026

This study compares three representative model families under identical data and validation protocols: logistic regression as the interpretable linear baseline, a multilayer perceptron as a deep learning representative, and gradient-boosted trees as a tree-ensemble representative. The data set holds 30,000 synthetic consumer lending records modeled after public peer-to-peer lending loan characteristics, with features that include the credit score, loan grade, interest rate, debt-to-income ratio, and revolving balance. The default rate is 34.0 percent. Every model is evaluated with stratified five-fold cross-validation across seven metrics, and differences are confirmed with the bootstrap DeLong and Friedman tests.

Three research questions are addressed. The first asks whether deep learning, in the form of a multilayer perceptron, significantly outperforms logistic regression for credit default prediction. The second asks whether gradient-boosted trees outperform both logistic regression and the multilayer perceptron. The third asks which model achieves the best probability calibration.

This paper makes four contributions. It provides a controlled multi-model comparison on a fixed credit-default data set under identical preprocessing and stratified five-fold cross-validation, which isolates the effect of model family from confounds such as feature engineering and split variance. It shows that the ranking of models reverses across metrics, since logistic regression leads on the area under the curve, precision, and calibration, while gradient-boosted trees lead on F1 and recall, and it explains the mechanism behind this reversal. It validates all pairwise differences in the area under the curve with the bootstrap DeLong test and the overall ranking with the Friedman test, and it separates statistical from practical significance at a sample size of 30,000. Finally, it translates the findings into deployment guidance for regulated lending, where interpretability and calibration often outweigh marginal gains in discrimination.

The gap addressed by this work is methodological as much as empirical. Many published comparisons vary the data set, the preprocessing, and the tuning budget at the same time as the model family, which makes the source of any reported gain hard to attribute [6], [8]. Reported improvements are also frequently summarized by one metric, even though a lender cares about discrimination, calibration, and the detection rate at once. By fixing the data, the preprocessing, and the validation protocol, and by reporting a full panel of metrics with significance tests, this study attributes any difference to the model family alone and exposes the trade-offs that a single number would hide.

II. Related Work

A. Traditional Credit Scoring

Logistic regression and scorecards remain dominant in industry practice because of the transparency and ease of governance these methods provide, and the Basel framework reinforces a preference for interpretable models [1]. Transparency, auditability, and explainability are now treated as first-class requirements for credit models rather than afterthoughts [2]. Systematic reviews of the field consistently find that, despite a proliferation of complex methods, simple classifiers stay remarkably competitive and no single technique dominates across data sets [5], [6], [9]. Recent reviews of consumer credit risk assessment reach the same conclusion and call for careful, metric-aware benchmarking [8].

B. Machine Learning for Credit Risk

Several benchmarks have measured machine learning approaches for credit-score prediction [4], and surveys have tracked the latest trends in credit-scoring methods [7]. One influential result shows that logistic regression augmented with non-linear decision-tree effects captures most of the achievable gain while keeping interpretability [10]. Empirical comparisons that place logistic regression next to random forests, support vector machines, and neural networks report only modest and data-dependent gains for the more complex models [11], [12]. Gradient boosting has become a strong contender in this space, with studies on personal default and bankruptcy prediction reporting competitive accuracy [13], [14]. Deep learning applications report mixed outcomes, and a direct study of whether deep learning helps credit scoring concluded that the gains over gradient boosting are limited [3], while the explainability of such models has become a parallel focus [15].

C. Deep Learning and Tabular Models

Gradient-boosted trees are the workhorse of structured prediction [16], and specialized tabular architectures such as attentive interpretable networks and gradient-boosting graph networks aim to close the gap with deep learning [17]-[19]. Deep models have nevertheless been applied directly to credit default, including penalized deep networks for feature extraction and deep learning on borrower-generated text in peer-to-peer lending [20], [21], and recent surveys map this fast-growing area [22]. Large benchmarks of deep learning on tabular data, however, repeatedly report that tree ensembles still match or beat deep networks on structured inputs [23]-[26], which supports the view that complexity does not automatically improve structured prediction. Beyond finance, deep architectures have advanced medical image analysis [27], [28], biomedical signal detection [29], and multimedia quality assessment for streaming and immersive applications [30], [31]. The present study examines whether such gains carry over to structured credit data.

D. Comparative Studies

A growing number of studies place traditional and modern models side by side on credit data. Comparative evaluations on credit-card and consumer portfolios report that ensemble and deep models can edge out a linear baseline, yet the margin is usually small and depends on the data set and the chosen metric [32]-[34]. Loan-default studies that apply boosting, random forests, and deep networks reach similar conclusions, with tree ensembles and boosted models often leading on default portfolios [35]-[37]. Further empirical comparisons on bank and online-lending portfolios, including studies that add alternative data sources and hybrid pipelines, confirm that gains over a strong baseline are modest and uneven [38]-[40], while additive gradient-boosting models recover much of the accuracy of opaque ensembles with greater transparency [41]. Broad reviews of artificial-intelligence methods for credit risk likewise find no

universally dominant technique and stress careful, like-for-like evaluation, which is the design principle adopted here [42].

E. Class Imbalance and Resampling

Because defaults are the minority class, imbalance handling is a recurring concern. Resampling studies and cost-sensitive learning report measurable effects on credit risk prediction [43], [44], and gradient boosting variants designed for imbalanced credit scoring extend this line of work [45]. Oversampling has been pushed further with explainability-aware synthetic sampling and generative adversarial oversampling [46], [47], and hybrid deep models paired with combined over- and under-sampling report gains on skewed credit data [48]. These methods change the decision threshold implicitly, which is one mechanism behind the precision-recall shifts observed in the present comparison.

F. Ensemble and Hybrid Architectures

Ensemble and hybrid models are a dominant theme in recent credit scoring. Interpretable ensembles and tree-based feature transformation aim to keep accuracy high while exposing the drivers of each decision [49], [50]. Feature-enhanced and voting-based hybrid ensembles add outlier handling and feature construction to the pipeline [51], [52], tree-based methods have been combined with deep learning for default prediction [53], and ensembles of explainable models have been assembled to balance performance with transparency [54]. A consistent message across these studies is that careful feature handling, rather than model family alone, drives most of the measurable gain.

G. Explainability, Fairness, and Evaluation

As credit models grow more complex, post-hoc explainability has become essential. Methods based on additive feature attribution have been used to interpret deep credit-scoring models [55], explainable models have been built specifically for consumer credit [56], and broader reviews survey explainable artificial intelligence in finance and inherently interpretable scoring models [57], [58]. Interpretable neural and gradient-boosting models, transparency-oriented frameworks, and decision-support studies extend this line to consumer and peer-to-peer lending [59]-[63]. Fairness has emerged as a parallel requirement, with formal assessments of fairness in credit scoring and its profit implications, and with frameworks that balance fairness against accuracy [64], [65]. Evaluation must also be metric-aware, since discrimination, calibration, and detection rate need not agree. Recent work surveys classifier calibration [66], warns that the area under the receiver operating characteristic curve can mislead on imbalanced data [67], and proposes cost-sensitive scoring metrics [68]. For comparing classifiers, this study adopts the DeLong test for correlated curves [69] and the Friedman test for ranking models across folds [70]. These tools frame the metric-aware comparison that follows.

Trust and fairness have become central to credit modeling. Explainable models built on boosting and additive attribution aim to make individual decisions auditable [71], [72], and studies that ask whether automated decisions can be trusted call for transparent, well-validated pipelines [73]. Fairness assessments measure disparate impact across protected groups and propose mitigation that trades a small accuracy cost for improved equity [74], [75]. Because a single metric can mislead under imbalance and asymmetric costs, recent work also proposes cost-sensitive and temporally aware evaluation frameworks [76] and threshold-free summary measures such as the H-measure [77]. The present study follows this guidance by reporting seven complementary metrics rather than a single headline figure.

III. METHOD

A. Preliminaries

The comparison covers three model families that span the complexity range used in credit scoring. Logistic regression is a generalized linear model that estimates the log-odds of default as a weighted sum of features, so its coefficients map directly to a scorecard and to adverse-action reasons, which makes it the regulatory reference model [1]. The multilayer perceptron is a feed-forward neural network that stacks non-linear transformations, so it can represent feature interactions that a linear model cannot, at the cost of transparency [24]. Gradient-boosted trees build an additive ensemble of shallow decision trees, each correcting the errors of the previous ones, and form the dominant method for structured prediction [16]. The central question is whether the added capacity of the second and third families translates into a consistent and well-characterized gain over the first on consumer credit data, once the preprocessing and the evaluation protocol are held fixed.

B. Data

The study uses 30,000 synthetic consumer lending records modeled after public peer-to-peer lending loan characteristics. The data generation builds in realistic correlations between loan grade, credit score, interest rate, and default probability. Table I summarizes the data set.

TABLE I. DATA SET DESCRIPTION

Property	Value
Samples	30,000 (synthetic)
Model features	22
Default rate	34.0%
Feature types	Numerical and encoded categorical
Borrower profile	Annual income, employment length, credit score
Loan terms	Loan amount, interest rate, installment, term
Credit behavior	Debt-to-income, revolving balance and use, open and total accounts
Delinquency	Delinquencies, recent inquiries, public records
Derived	Loan-to-income, installment-to-income, credit use
Encoded categorical	Grade, home ownership, purpose, verification

The default probability is generated through a logistic function of loan grade, credit score, debt-to-income ratio, delinquency history, and interest rate, which produces a realistic 34.0 percent default rate with grade-correlated risk.

All features pass through a single shared preparation pipeline, so every model receives identical inputs.

Continuous variables are standardized to zero mean and unit variance, which is required for stable training of the logistic regression and the multilayer perceptron and is harmless for the gradient-boosted trees. Categorical fields, such as the loan grade, the home-ownership status, the loan purpose, and the income-verification status, are integer-encoded. Three ratio features, namely the loan-to-income ratio, the installment-to-income ratio, and a credit-use ratio, are derived from the raw fields to expose relationships that a linear model cannot recover on its own. No feature selection is applied, so all 22 features enter every model. This shared preparation isolates the effect of the model family, which is the central aim of the comparison.

C. Models

Logistic regression uses L2 regularization with a regularization constant of 1.0 and a maximum of 1,000 iterations. All features are standardized to zero mean and unit variance.

The multilayer perceptron uses an architecture of 128, 64, and 32 hidden units with rectified linear activation, batch normalization after the first two layers, dropout rates of 0.3, 0.3, and 0.2, and the Adam optimizer with a learning rate of 0.001, a batch size of 256, and up to 50 epochs with early stopping at a patience of five. The sigmoid output is given in Equation (1):

$$\hat{y} = \sigma(\mathbf{W}_3\, g(\mathrm{BN}(\mathbf{W}_2\, g(\mathrm{BN}(\mathbf{W}_1 \mathbf{x}))))) \quad (1)$$

where g of a value equals the maximum of zero and that value, which denotes the rectified linear unit, and BN denotes batch normalization.

The gradient-boosted trees model uses 500 estimators, a maximum depth of six, a learning rate of 0.05, a subsample ratio of 0.8, a column subsample ratio of 0.8, and class-weight balancing through a scaled positive weight [16].

D. Evaluation Metrics

Seven complementary metrics are reported, because no single metric captures every aspect of a credit model. Let the counts of true positives, false positives, true negatives, and false negatives at the classification threshold be written as TP, FP, TN, and FN. Precision, recall, and the F1-score are defined in Equations (2) to (4).

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (3)$$

$$\mathrm{F1} = \frac{2\,\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (4)$$

Precision measures how many flagged borrowers truly default, while recall measures how many true defaults are caught, and the F1-score is the harmonic mean of precision and recall. The area under the receiver operating characteristic curve measures ranking quality across all thresholds, and the average precision summarizes the precision-recall curve. Probability calibration is measured by the Brier score in Equation (5), where a lower value is better.

$$\mathrm{BS} = \tfrac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2 \quad (5)$$

where the predicted default probability is written as the estimated p, the binary outcome as y, and the number of records as N. Discrimination, calibration, and detection rate need not agree, so reporting all seven metrics is essential for a fair comparison.

E. Evaluation Protocol

Stratified five-fold cross-validation preserves the 34.0 percent default rate in each fold. The reported metrics are the area under the receiver operating characteristic curve, the F1-score, precision, recall, accuracy, average precision, and the Brier score. Two statistical tests are applied. The bootstrap DeLong test is a non-parametric comparison of correlated curves through 2,000 bootstrap replications [69]. The Friedman test is a non-parametric test for ranking the three models across the five folds [70]. The end-to-end procedure is summarized in Algorithm 1. It standardizes features, partitions the data into five stratified folds, trains each model on four folds and predicts on the held-out fold, aggregates the seven metrics as a mean across folds, and then applies the two tests.

Algorithm 1 Comparative Credit Default Modeling Pipeline

Require: data set D = {(x_i, y_i)}; models {logistic, perceptron, boosted}; folds K = 5
Ensure: per-model metric estimates and significance tests
1: Standardize features to zero mean and unit variance
2: Partition D into K stratified folds preserving the 34.0% default rate
3: for each model m do
for k = 1 to K do
Train m on all folds except k; predict probabilities on fold k
Compute AUC, F1, precision, recall, accuracy, AP, and Brier
end for
Aggregate each metric as the mean over the K folds
4: end for
5: Compare AUC curves pairwise with the bootstrap DeLong test (B = 2000)
6: Rank the models across folds with the Friedman test
7: return aggregated metrics and test statistics

F. Reproducibility and Validation Design

The comparison is designed so that the only variable across the three models is the model family. The same feature matrix, the same stratified five-fold partition, and the same random seed are used for every model, and no model-specific feature engineering is applied. Hyperparameters are fixed in advance from common defaults rather than tuned on the test folds, which avoids optimistic bias but also means the reported figures are lower bounds on what tuning could achieve. Each metric is aggregated as the mean across the five held-out folds, and the standard deviation across folds is reported for the area under the curve to show stability. The bootstrap DeLong test reuses the same fold predictions, so the significance test reflects the same data the metrics

summarize. This fixed-design protocol is what allows any observed difference to be attributed to the model family rather than to preprocessing or split variance, and it makes the pipeline straightforward to reproduce from the stated settings.

IV. Results and Discussion

A. Model Performance

Table II reports the five-fold cross-validation results. Logistic regression reaches the highest area under the curve, at 0.726, followed closely by the multilayer perceptron at 0.723 and then the gradient-boosted trees at 0.705. The metric rankings are not uniform. The gradient-boosted model reaches the highest F1-score, at 0.548, and the highest recall, at 0.584, so it detects far more defaults than logistic regression, whose recall is 0.365, or the multilayer perceptron, whose recall is 0.378, but it does so at the cost of lower precision, at 0.516 against 0.624.

TABLE II. Five-Fold Cross-Validation Results

Metric	Logistic	Perceptron	Boosted
AUC-ROC	**0.726**	0.723	0.705
Brier score	**0.192**	0.193	0.211
F1-score	0.461	0.470	**0.548**
Precision	**0.624**	0.620	0.516
Recall	0.365	0.378	**0.584**
Accuracy	0.709	0.710	0.672
Avg. precision	**0.580**	0.577	0.552

Figure 1 visualizes the comparison across the area under the curve, the F1-score, and average precision. Logistic regression and the multilayer perceptron show similar profiles, with high precision and low recall, while the gradient-boosted model trades precision for recall.

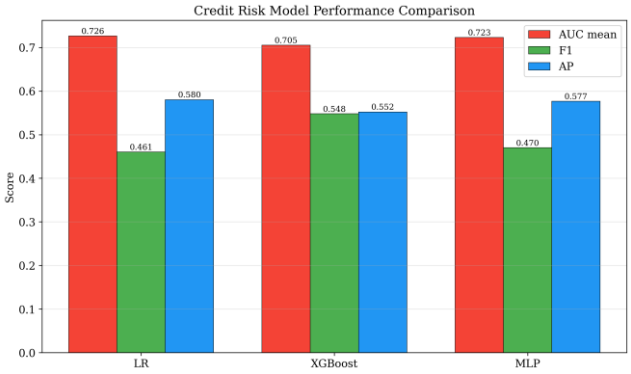

Fig. 1. Model performance across the area under the curve, the F1-score, and average precision. Logistic regression and the multilayer perceptron cluster together, while the gradient-boosted model shows a different precision-recall trade-off.

Figure 2 presents the receiver operating characteristic curves for all three models. The curves for logistic regression and the multilayer perceptron nearly overlap, while the gradient-boosted model shows lower discrimination across most operating points.

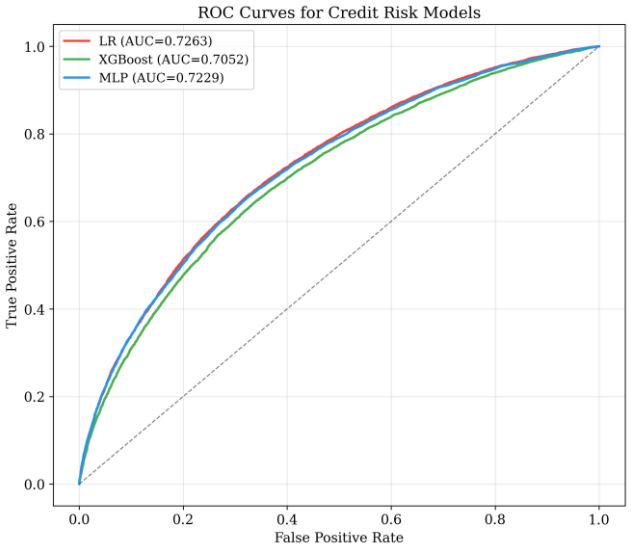

Fig. 2. Receiver operating characteristic curves for logistic regression (area 0.726), the multilayer perceptron (area 0.723), and the gradient-boosted model (area 0.705). The first two curves are nearly indistinguishable.

B. Statistical Tests

Table III presents the statistical test results. All pairwise bootstrap DeLong tests are significant, with every p-value below 0.001, which confirms that the differences in the area under the curve are not due to sampling variability. The difference between logistic regression and the multilayer perceptron is 0.004, with a 95 percent confidence interval from 0.002 to 0.005, which is statistically significant but practically small. The difference between logistic regression and the gradient-boosted model is larger, at 0.021, with an interval from 0.018 to 0.024. The Friedman test confirms that the overall ranking is significant, with a chi-square statistic of 10.0 and a p-value of 0.007.

TABLE III. Pairwise DeLong and Friedman Test Results

Comparison	Diff.	z	p	95% CI
Logistic vs. perceptron	+0.004	5.74	<0.001	[0.002, 0.005]
Logistic vs. boosted	+0.021	13.56	<0.001	[0.018, 0.024]
Perceptron vs. boosted	+0.018	11.51	<0.001	[0.015, 0.021]
Friedman chi-square	10.0		0.007	

Figure 3 shows the stability of the area under the curve across folds. All three models are consistent, and logistic regression keeps a slight edge in every fold.

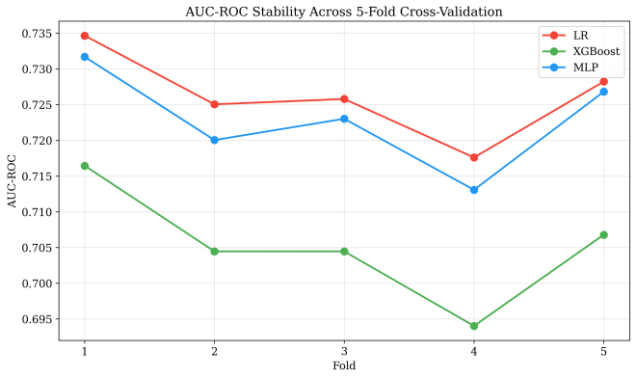

Fig. 3. Area under the curve across the five cross-validation folds. Logistic regression consistently reaches the highest value in each fold, with the multilayer perceptron close behind and the gradient-boosted model lower.

C. *Feature Importance*

Table IV reports the top ten features by the gain-based importance of the gradient-boosted model. The encoded loan grade dominates, at 24.6 percent, which is consistent with its direct encoding of lender-assessed creditworthiness. The credit score ranks second, at 5.4 percent, followed by the interest rate, at 4.5 percent, and the installment-to-income ratio, at 3.8 percent.

TABLE IV. TOP TEN FEATURES BY GAIN-BASED IMPORTANCE

Feature	Importance (%)	Rank
Grade (encoded)	24.6	1
Credit score	5.4	2
Interest rate	4.5	3
Installment-to-income	3.8	4
Debt-to-income	3.8	5
Installment	3.7	6
Loan-to-income	3.7	7
Revolving balance	3.6	8
Annual income	3.5	9
Revolving use	3.5	10

Figure 4 visualizes the top features. The concentration of importance in the encoded grade suggests that this feature already captures much of the information also held by the credit score and the interest rate, since the loan grade is typically assigned on the basis of creditworthiness.

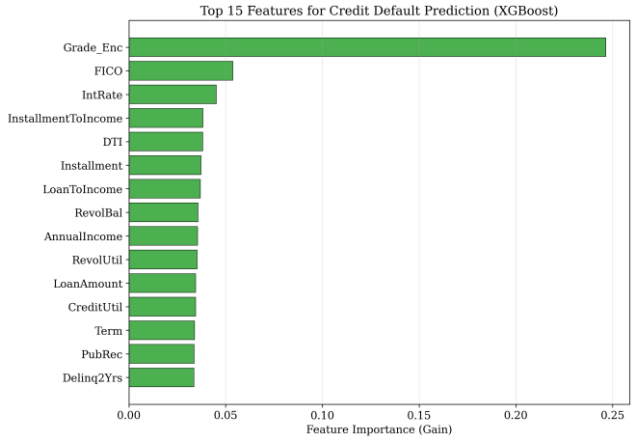

Fig. 4. Top features by gain-based importance for the gradient-boosted model. The loan grade dominates at 24.6 percent, while the remaining features each contribute between three and five percent.

D. *Operating-Point Analysis*

The headline metrics are read at each model native decision threshold, so the precision-recall contrast in Table II reflects where each model places that threshold. At the default threshold of 0.5, logistic regression classifies few applicants as defaults, which yields high precision, at 0.624, but low recall, at 0.365. The class-weighted gradient-boosted model places its boundary far lower, which lifts recall to 0.584 while precision falls to 0.516. The multilayer perceptron sits close to logistic regression, with precision 0.620 and recall 0.378. Because the area under the curve and the average precision are threshold-free, the lead of logistic regression on those two metrics holds across all operating points, whereas the F1 and recall lead of the gradient-boosted model is a property of its lower native threshold rather than of better ranking. A lender that re-tuned every threshold to a common target recall would therefore narrow the F1 gap while preserving the ranking on the area under the curve, which is the more fundamental measure of separation.

E. *Error Analysis*

The errors split along the same precision-recall axis. At its native threshold logistic regression misses roughly two thirds of true defaults, since recall is 0.365, but most of the applicants it does flag are genuine defaults, since precision is 0.624. The gradient-boosted model inverts this profile, catching 0.584 of defaults at the cost of a larger share of false alarms, since precision is 0.516. In monetary terms, the first profile minimizes the rejection of good borrowers while accepting more undetected defaults, and the second minimizes undetected defaults while rejecting more good borrowers. The calibration figures sharpen this picture: the Brier score of 0.192 for logistic regression against 0.211 for the gradient-boosted model means the predicted probabilities of the linear model are closer to observed default frequencies, so its scores can be used directly for pricing and provisioning, while the scores of the gradient-boosted model would need recalibration before such use.

F. *The Metric-Dependent Trade-Off*

A key finding is that the model ranking depends on the evaluation metric. Logistic regression ranks first on the area under the curve, precision, and calibration, while the gradient-boosted model ranks first on F1 and recall. This happens because the class-weight balancing of the gradient-boosted model encourages higher recall at the cost of precision, which shifts the decision boundary toward classifying more applicants as defaults. The default threshold of 0.5 used by logistic regression produces a conservative classifier with high precision but low recall.

This trade-off has practical meaning. When the priority is to identify all potential defaults, for example in risk-averse lending, the higher recall of the gradient-boosted model, at 0.584 against 0.365, is preferable. When the priority is to limit false alarms, for example to avoid rejecting good borrowers, the higher precision of logistic regression, at 0.624 against 0.516, is superior. The right choice is therefore a function of the cost structure of the lender, not of model sophistication alone [68].

G. Why Logistic Regression Reaches the Highest Discrimination

The advantage of logistic regression on the area under the curve is consistent with the flat maximum effect in credit scoring [1]. When the relationships between features and default are close to monotonic, where a higher credit score and a higher grade both lower risk, added model complexity mostly adds noise. The logistic generation process of the synthetic data produces relationships that are well suited to logistic regression, and the result aligns with recent tabular benchmarks in which added complexity rarely helps structured data [23], [24], [26].

H. Statistical against Practical Significance

The difference of 0.004 in the area under the curve between logistic regression and the multilayer perceptron is statistically significant, with a p-value below 0.001, but practically negligible. At a sample size of 30,000, even tiny differences reach significance through the bootstrap DeLong test. The confidence interval from 0.002 to 0.005 confirms that the difference is real but small. Financial institutions should weigh this marginal advantage against the computational cost and reduced interpretability of the multilayer perceptron [2].

I. Generalizability and Robustness

Two properties support cautious generalization of the ranking, and one limits it. The first supporting property is fold stability: the area under the curve varies by only about 0.006 across folds for each model, and logistic regression keeps its lead in every fold, so the ordering is not an artifact of a single split. The second is mechanism: the lead of the linear model is explained by the near-monotonic, grade-dominated structure of the data, a structure common to real credit portfolios, rather than by a chance fit. The limiting property is scope: the data are synthetic and drawn from one generation process, so the absolute numbers should not be read as production performance, only the relative ordering under a controlled design. The literature reviewed above reports the same qualitative pattern across many real data sets, which raises confidence that the metric-dependent reversal, rather than the exact values, is the transferable finding.

J. Comparison with Prior Work

The metric-dependent reversal observed here aligns with the broader literature. Comparative studies on credit and loan data report that deep and ensemble models lead only by small margins that depend on the data set and the metric [32]-[35]. The result that a linear model leads on discrimination and calibration while a boosted model leads on recall is consistent with reviews that find no single technique dominant across credit portfolios [42] and with the flat-maximum effect long noted in credit scoring [1]. The practical message echoes the trust and fairness literature, since a small, unexplained gain rarely justifies the added governance burden in regulated lending [73], [74].

K. Threats to Validity

Several factors bound these conclusions. Construct validity rests on the synthetic data, which reproduce realistic correlations but cannot capture every non-linear pattern in a real portfolio, and the logistic generation process may modestly favor the linear model. Internal validity is supported by identical preprocessing, a fixed stratified protocol, and significance testing, yet the single multilayer-perceptron configuration was not exhaustively tuned, so its ceiling may be underestimated. External validity is limited by the single data set and the absence of temporal or macroeconomic drift, so cost-sensitive and temporally aware evaluation would strengthen transfer to production conditions [76].

L. Limitations

Two further limitations remain beyond the threats noted above. No cost-sensitive evaluation is performed, although different misclassification costs are critical in lending [64]. Class-weight balancing was applied to the gradient-boosted model but not to logistic regression or the multilayer perceptron, which partly explains the recall difference.

M. A Decision-Theoretic View

The metric-dependent result is clearest through a decision-theoretic lens. A lending decision has two error costs: the loss from approving a borrower who later defaults, and the foregone margin from declining a borrower who would have repaid. The optimal threshold sets the predicted default probability at the point where the expected cost of approval equals the expected margin, which depends on the loss-given-default and the price of the loan, not on the model alone [68]. Under this view, the area under the curve and the Brier score describe how well a model orders and calibrates risk, which are properties the lender needs before any threshold is chosen, while F1 and recall describe one particular threshold. The calibrated linear model therefore supplies the inputs a cost-sensitive policy requires, and the threshold can then be moved to reach any target recall. Reporting performance at a single fixed threshold, as much of the comparative literature does, conflates the quality of the model with the choice of operating point, which a temporally aware, cost-sensitive evaluation would separate [76].

N. Deployment Considerations

Beyond raw performance, the choice of model in regulated lending depends on operational factors. Logistic regression yields a small, fixed set of coefficients that map directly to a scorecard, which supports adverse-action reasons, ongoing monitoring, and independent validation at low cost [2]. The multilayer perceptron and the gradient-boosted model need post-hoc explanation tools to meet the same transparency bar, which adds engineering and governance overhead [57]. The gradient-boosted model is attractive when the goal is to catch as many defaults as possible, because its higher recall directly reduces missed losses, but its lower precision raises the number of good applicants who are wrongly declined. A cost-sensitive threshold, set from the expected loss of a missed default weighed against the lost margin on a rejected good borrower, is therefore more useful than a fixed threshold of 0.5 [68]. In practice, a staged design that uses logistic regression for the primary, explainable decision and a gradient-boosted model as a secondary recall-oriented screen can combine the strengths of both families.

O. Implications for Practice and Research

Two implications follow for practice. First, model selection should begin with the business objective and its cost structure, then fix the metric, and only then choose the model, rather than the reverse order. A lender that prices missed defaults heavily should favor the recall-oriented boosted model, while a lender that must justify every decline should favor the calibrated linear model [68]. Second, governance effort should be budgeted alongside accuracy, because the marginal discrimination gain from added complexity is small on structured credit data and may not survive independent validation [2]. For research, the findings argue for reporting full metric panels with significance tests and for testing whether conclusions hold on real, temporally split portfolios rather than on a single static sample [76].

V. CONCLUSION

Logistic regression reaches the highest area under the receiver operating characteristic curve, at 0.726, and the best calibration, with a Brier score of 0.192, for credit default prediction, ahead of the multilayer perceptron at 0.723 and the gradient-boosted model at 0.705, with all pairwise differences significant. The gradient-boosted model, however, reaches the highest F1-score, at 0.548, and the highest recall, at 0.584, detecting 58 percent of defaults against 37 percent for logistic regression. The Friedman test confirms significant differences in the model ranking, with a chi-square statistic of 10.0 and a p-value of 0.007. These results show that model selection for credit scoring should be metric-driven. Evaluation focused on the area under the curve favors logistic regression, while evaluation focused on recall favors the gradient-boosted model. For structured credit data, increased model complexity does not uniformly improve all performance metrics. Future work will extend the comparison to real lending portfolios, cost-sensitive objectives, and model selection that is constrained by explainability [22], [58].

REFERENCES

[1] L. Thomas, J. Crook, and D. Edelman, "Credit Scoring and Its Applications, Second Edition," 2017. doi: 10.1137/1.9781611974560.

[2] M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek, "Transparency, auditability, and explainability of machine learning models in credit scoring," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 70-90, 2021. doi: 10.1080/01605682.2021.1922098.

[3] B. R. Gunnarsson, S. vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, "Deep learning for credit scoring: Do or don't?," *European Journal of Operational Research*, vol. 295, no. 1, pp. 292-305, 2021. doi: 10.1016/j.ejor.2021.03.006.

[4] V. Moscato, A. Picariello, and G. Sperlí, "A benchmark of machine learning approaches for credit score prediction," *Expert Systems with Applications*, vol. 165, pp. 113986, 2021. doi: 10.1016/j.eswa.2020.113986.

[5] H. Ayari, P. R. Guetari, and P. N. Kraïem, "Machine learning powered financial credit scoring: a systematic literature review," *Artificial Intelligence Review*, vol. 59, no. 1, art. no. 13, 2025, doi: 10.1007/s10462-025-11416-2.

[6] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Applied Soft Computing*, vol. 91, pp. 106263, 2020. doi: 10.1016/j.asoc.2020.106263.

[7] A. Markov, Z. Seleznyova, and V. Lapshin, "Credit scoring methods: Latest trends and points to consider," *The Journal of Finance and Data Science*, vol. 8, pp. 180-201, 2022. doi: 10.1016/j.jfds.2022.07.002.

[8] X. Zhang and L. Yu, "Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods," *Expert Systems with Applications*, vol. 237, pp. 121484, 2024, doi: 10.1016/j.eswa.2023.121484.

[9] J. P. Noriega, L. A. Rivera, and J. A. Herrera, "Machine Learning for Credit Risk Prediction: A Systematic Literature Review," *Data*, vol. 8, no. 11, pp. 169, 2023. doi: 10.3390/data8110169.

[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*, vol. 297, no. 3, pp. 1178-1192, 2022. doi: 10.1016/j.ejor.2021.06.053.

[11] S. Mestiri, "Credit scoring using machine learning and deep Learning-Based models," *Data Science in Finance and Economics*, vol. 4, no. 2, pp. 236-248, 2024, doi: 10.3934/dsfe.2024009.

[12] V. Chang, S. Sivakulasingam, H. Wang, S. T. Wong, M. A. Ganatra, and J. Luo, "Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers," *Risks*, vol. 12, no. 11, pp. 174, 2024, doi: 10.3390/risks12110174.

[13] N. Nguyen and D. Ngo, "Comparative analysis of boosting algorithms for predicting personal default," *Cogent Economics & Finance*, vol. 13, no. 1, art. no. 2465971, 2025, doi: 10.1080/23322039.2025.2465971.

[14] S. Ben Jabeur, N. Stef, and P. Carmona, "Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering," *Computational Economics*, vol. 61, no. 2, pp. 715-741, 2022, doi: 10.1007/s10614-021-10227-1.

[15] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," *Computational Economics*, vol. 57, no. 1, pp. 203-216, 2020. doi: 10.1007/s10614-020-10042-0.

[16] T. Chen, and C. Guestrin, "XGBoost," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016. doi: 10.1145/2939672.2939785.

[17] S. Ö. Arik, and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679-6687, 2021. doi: 10.1609/aaai.v35i8.16826.

[18] C. Y. Hsu, and C. T. Li, "GTab: Gradient Boosting Bipartite Graph Neural Networks for Holistic Tabular Data Predictions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 38, no. 6, pp. 3669-3683, 2026. doi: 10.1109/tkde.2026.3678153.

[19] S. Wang and X. Zhang, "Research on Credit Default Prediction Model Based on TabNet-Stacking," *Entropy*, vol. 26, no. 10, pp. 861, 2024, doi: 10.3390/e26100861.

[20] C. Lin, N. Qiao, W. Zhang, Y. Li, and S. Ma, "Default risk prediction and feature extraction using a penalized deep neural network," *Statistics and Computing*, vol. 32, no. 5, art. no. 76, 2022, doi: 10.1007/s11222-022-10140-z.

[21] J. Kriebel and L. Stitz, "Credit default prediction from user-generated text in peer-to-peer lending using deep learning," *European Journal of Operational Research*, vol. 302, no. 1, pp. 309-323, 2022, doi: 10.1016/j.ejor.2021.12.024.

[22] I. D. Mienye, E. Esenogho, and C. Modisane, "Deep Learning for Credit Risk Prediction: A Survey of Methods, Applications, and Challenges," *Information*, vol. 17, no. 4, pp. 395, 2026, doi: 10.3390/info17040395.

[23] R. Shwartz-Ziv, and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84-90, 2022. doi: 10.1016/j.inffus.2021.11.011.

[24] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7499-7519, 2024, doi: 10.1109/tnnls.2022.3229161.

[25] S. Somvanshi, S. Das, S. Javed, G. Antariksa, and A. Hossain, "A Survey on Tabular Data: From Tree-based Methods to Tabular Deep Learning," *ACM Computing Surveys*, art. no. 3807777, 2026, doi: 10.1145/3807777.

[26] A. Shmuel, O. Glickman, and T. Lazebnik, "A comprehensive benchmark of machine and deep learning models on structured data for regression and classification," *Neurocomputing*, vol. 655, pp. 131337, 2025, doi: 10.1016/j.neucom.2025.131337.

[27] A. A. Laghari, V. V. Estrela, and S. Yin, "How to Collect and Interpret Medical Pictures Captured in Highly Challenging Environments that Range from Nanoscale to Hyperspectral Imaging," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 20, art. e281222212228, 2024. doi: 10.2174/1573405619666221228094228.

[28] S. Yin et al., "Brain CT image classification based on mask RCNN and attention mechanism," *Scientific Reports*, vol. 14, no. 1, art. 29300, 2024. doi: 10.1038/s41598-024-78566-1.

[29] A. A. Laghari, Y. Sun, M. Alhussein, K. Aurangzeb, M. S. Anwar, and M. Rashid, "Deep residual-dense network based on bidirectional recurrent neural network for atrial fibrillation detection," *Scientific Reports*, vol. 13, no. 1, art. 15109, 2023. doi: 10.1038/s41598-023-40343-x.

[30] A. A. Laghari et al., "The state of art and review on video streaming," *Journal of High Speed Networks*, vol. 29, no. 3, pp. 211-236, 2023. doi: 10.3233/jhs-222087.

[31] A. A. Laghari et al., "Quality of experience assessment in virtual/augmented reality serious games for healthcare: A systematic literature review," *Technology and Disability*, vol. 36, no. 1-2, pp. 17-28, 2024. doi: 10.3233/tad-230035.

[32] N. T. H. Thuy, N. T. V. Ha, N. N. Trung, V. T. T. Binh, N. T. Hang, and V. T. Binh, "Comparing the Effectiveness of Machine Learning and Deep Learning Models in Student Credit Scoring: A Case Study in Vietnam," *Risks*, vol. 13, no. 5, pp. 99, 2025, doi: 10.3390/risks13050099.

[33] I. Mapfumo and T. Shongwe, "Performance Evaluation of Machine Learning and Deep Learning Models for Credit Risk Prediction," *Journal of Risk and Financial Management*, vol. 19, no. 3, pp. 210, 2026, doi: 10.3390/jrfm19030210.

[34] R. Bhandary and B. K. Ghosh, "Credit Card Default Prediction: An Empirical Analysis on Predictive Performance Using Statistical and Machine Learning Methods," *Journal of Risk and Financial Management*, vol. 18, no. 1, pp. 23, 2025, doi: 10.3390/jrfm18010023.

[35] Y. Song, Y. Wang, X. Ye, R. Zaretzki, and C. Liu, "Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme," *Information Sciences*, vol. 629, pp. 599-617, 2023, doi: 10.1016/j.ins.2023.02.014.

[36] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, "Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction," *Mathematics*, vol. 12, no. 21, pp. 3423, 2024, doi: 10.3390/math12213423.

[37] X. Zhang, T. Zhang, L. Hou, X. Liu, Z. Guo, Y. Tian, and Y. Liu, "Data-Driven Loan Default Prediction: A Machine Learning Approach for Enhancing Business Process Management," *Systems*, vol. 13, no. 7, pp. 581, 2025, doi: 10.3390/systems13070581.

[38] N. Suhadolnik, J. Ueyama, and S. Da Silva, "Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach," *Journal of Risk and Financial Management*, vol. 16, no. 12, pp. 496, 2023, doi: 10.3390/jrfm16120496.

[39] T. Berhane, T. Melese, and A. M. Seid, "Performance Evaluation of Hybrid Machine Learning Algorithms for Online Lending Credit Risk Prediction," *Applied Artificial Intelligence*, vol. 38, no. 1, art. no. 2358661, 2024, doi: 10.1080/08839514.2024.2358661.

[40] R. Hlongwane, K. K. K. M. Ramaboa, and W. Mongwe, "Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data," *PLOS ONE*, vol. 19, no. 5, pp. e0303566, 2024, doi: 10.1371/journal.pone.0303566.

[41] Y. Zou, M. Xia, and X. Lan, "Interpretable credit scoring based on an additive extreme gradient boosting," *Chaos, Solitons & Fractals*, vol. 194, pp. 116216, 2025, doi: 10.1016/j.chaos.2025.116216.

[42] V. Amarnadh and N. R. Moparthi, "Comprehensive review of different artificial intelligence-based methods for credit risk assessment in data science," *Intelligent Decision Technologies*, vol. 17, no. 4, pp. 1265-1282, 2023, doi: 10.3233/idt-230190.

[43] Z. Zhao, T. Cui, S. Ding, J. Li, and A. G. Bellotti, "Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction," *Mathematics*, vol. 12, no. 5, pp. 701, 2024, doi: 10.3390/math12050701.

[44] Y. Chen, R. Calabrese, and B. Martin-Barragan, "Interpretable machine learning for imbalanced credit scoring datasets," *European Journal of Operational Research*, vol. 312, no. 1, pp. 357-372, 2024, doi: 10.1016/j.ejor.2023.06.036.

[45] J. Zhang, R. Calabrese, and Y. Dong, "A novel generalised extreme value gradient boosting decision tree for the class imbalanced problem in credit scoring," *Journal of the Operational Research Society*, vol. 76, no. 8, pp. 1513-1530, 2024, doi: 10.1080/01605682.2024.2418882.

[46] S. Han, H. Jung, P. D. Yoo, A. Provetti, and A. Cali, "NOTE: non-parametric oversampling technique for explainable credit scoring," *Scientific Reports*, vol. 14, no. 1, art. no. 26070, 2024, doi: 10.1038/s41598-024-78055-5.

[47] I. N. M. Adiputra, P. C. Lin, and P. Wanchai, "The Effectiveness of Generative Adversarial Network-Based Oversampling Methods for Imbalanced Multi-Class Credit Score Classification," *Electronics*, vol. 14, no. 4, pp. 697, 2025, doi: 10.3390/electronics14040697.

[48] I. Aruleba and Y. Sun, "Enhanced credit risk prediction using deep learning and SMOTE-ENN resampling," *Machine Learning with Applications*, vol. 21, pp. 100692, 2025, doi: 10.1016/j.mlwa.2025.100692.

[49] Y. Liu, F. Huang, L. Ma, Q. Zeng, and J. Shi, "Credit scoring prediction leveraging interpretable ensemble learning," *Journal of Forecasting*, vol. 43, no. 2, pp. 286-308, 2023, doi: 10.1002/for.3033.

[50] J. Liu, J. Liu, C. Wu, and S. Wang, "Enhancing credit risk prediction based on ensemble tree-based feature transformation and logistic regression," *Journal of Forecasting*, vol. 43, no. 2, pp. 429-455, 2023, doi: 10.1002/for.3040.

[51] D. Yang, B. Xiao, M. Cao, and H. Shen, "A new hybrid credit scoring ensemble model with feature enhancement and soft voting weight optimization," *Expert Systems with Applications*, vol. 238, pp. 122101, 2024, doi: 10.1016/j.eswa.2023.122101.

[52] W. Zhang, D. Yang, and S. Zhang, "A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring," *Expert Systems with Applications*, vol. 174, pp. 114744, 2021, doi: 10.1016/j.eswa.2021.114744.

[53] H. He and Y. Fan, "A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction," *Expert Systems with Applications*, vol. 176, pp. 114899, 2021, doi: 10.1016/j.eswa.2021.114899.

[54] X. Sun, J. Liu, and Y. Zhang, "Enhancing Credit Risk Prediction through an Ensemble of Explainable Model," *Journal of Systems Science and Systems Engineering*, vol. 34, no. 5, pp. 619-640, 2025, doi: 10.1007/s11518-025-5663-y.

[55] L. O. Hjelkrem, and P. E. d. Lange, "Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data," *Journal of Risk and Financial Management*, vol. 16, no. 4, pp. 221, 2023. doi: 10.3390/jrfm16040221.

[56] D. U. Chacon, S. Lee, and J. Park, "An explainable machine learning model for consumer credit scoring in Mexico," *Emerging Markets Review*, vol. 71, pp. 101424, 2026. doi: 10.1016/j.ememar.2025.101424.

[57] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," *Artificial Intelligence Review*, vol. 57, no. 8, art. no. 216, 2024, doi: 10.1007/s10462-024-10854-8.

[58] J. Tu and Z. Wu, "Inherently interpretable machine learning for credit scoring: Optimal classification tree with hyperplane splits," *European Journal of Operational Research*, vol. 322, no. 2, pp. 647-664, 2025, doi: 10.1016/j.ejor.2024.10.046.

[59] F. M. Talaat, A. Aljadani, M. Badawy, and M. Elhosseini, "Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction," *Neural Computing and Applications*, vol. 36, no. 9, pp. 4847-4865, 2023, doi: 10.1007/s00521-023-09232-2.

[60] R. Hlongwane, K. Ramabao, and W. Mongwe, "A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards," *PLOS ONE*, vol. 19, no. 8, pp. e0308718, 2024, doi: 10.1371/journal.pone.0308718.

[61] M. K. Nallakaruppan, H. Chaturvedi, V. Grover, B. Balusamy, P. Jaraut, J. Bahadur, V. P. Meena, and I. A. Hameed, "Credit Risk Assessment and Financial Decision Support Using Explainable Artificial Intelligence," *Risks*, vol. 12, no. 10, pp. 164, 2024, doi: 10.3390/risks12100164.

[62] L. H. Li, A. K. Sharma, and S. T. Cheng, "Explainable AI based LightGBM prediction model to predict default borrower in social lending platform," *Intelligent Systems with Applications*, vol. 26, pp. 200514, 2025, doi: 10.1016/j.iswa.2025.200514.

[63] M. Atef, S. Ouf, W. Seoud, and M. I. Gabr, "A novel approach using explainable prediction of default risk in peer-to-peer lending based on machine learning models," *Neural Computing and Applications*, vol. 37, no. 26, pp. 21783-21803, 2025, doi: 10.1007/s00521-025-11489-8.

[64] 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.

[65] E. Kpatcha, "Balancing Fairness and Accuracy in Machine Learning-Based Probability of Default Modeling via Threshold Optimization," *Journal of Risk and Financial Management*, vol. 18, no. 12, pp. 724, 2025, doi: 10.3390/jrfm18120724.

[66] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, no. 9, pp. 3211-3260, 2023, doi: 10.1007/s10994-023-06336-7.

[67] M. Imani, M. Joudaki, A. Bagheri, and H. R. Arabnia, "Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-Score, H-Measure, and AUC-Based Metrics Across Diverse Classifiers," *Technologies*, vol. 14, no. 1, pp. 54, 2026, doi: 10.3390/technologies14010054.

[68] N. Khalili and M. A. Rastegar, "Optimal cost-sensitive credit scoring using a new hybrid performance metric," *Expert Systems with Applications*, vol. 213, pp. 119232, 2023, doi: 10.1016/j.eswa.2022.119232.

[69] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, vol. 44, no. 3, pp. 837, 1988. doi: 10.2307/2531595.

[70] M. Friedman, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675-701, 1937. doi: 10.1080/01621459.1937.10503522.

[71] L. Monje, R. A. Carrasco, and M. Sánchez-Montañés, "Machine Learning XAI for Early Loan Default Prediction," *Computational Economics*, vol. 67, no. 5, pp. 4033-4062, 2025, doi: 10.1007/s10614-025-10962-9.

[72] C. N. Nwafor, O. Nwafor, and S. Brahma, "Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach," *Scientific Reports*, vol. 14, no. 1, art. no. 25174, 2024, doi: 10.1038/s41598-024-75026-8.

[73] A. Alonso-Robisco and J. M. Carbó, "Should We Trust the Credit Decisions Provided by Machine Learning Models?," *Computational Economics*, vol. 66, no. 5, pp. 4245-4274, 2025, doi: 10.1007/s10614-025-10855-x.

[74] G. Babaei and P. Giudici, "How fair is machine learning in credit lending?," *Quality and Reliability Engineering International*, vol. 40, no. 6, pp. 3452-3464, 2024, doi: 10.1002/qre.3579.

[75] J. Lin and L. Yang, "Mitigating algorithmic bias in credit scoring support systems through adversarial learning," *Decision Support Systems*, vol. 205, pp. 114653, 2026, doi: 10.1016/j.dss.2026.114653.

[76] T. Sodnomdavaa and M. Sandagsuren, "Temporal and Cost-Sensitive Evaluation Framework for Credit Risk Modeling Under Distributional Shifts," *Risks*, vol. 14, no. 4, pp. 95, 2026, doi: 10.3390/risks14040095.

[77] D. J. Hand and C. Anagnostopoulos, "Notes on the H-measure of classifier performance," *Advances in Data Analysis and Classification*, vol. 17, no. 1, pp. 109-124, 2022, doi: 10.1007/s11634-021-00490-3.