Vol. 19 (2026): Proceedings of 2026 International Conference on Financial Technology, Mathematical Statistics, and Business Analytics (ICFMB 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/icfmb.202619396401

A Comparative Study of Traditional Statistical Models and Machine Learning Algorithms in Credit Risk Assessment

Hanrun Jin*

Stony Brook University, New York, United States

**Corresponding author: Hanrun Jin.*

Abstract

Credit risk assessment underpins lending decisions, pricing strategies, portfolio management, and regulatory capital allocation within modern financial systems. Logistic regression has historically served as the dominant modeling framework in credit scoring due to its probabilistic coherence and interpretability. In recent years, advances in machine learning—particularly tree-based ensemble methods such as Random Forest and Gradient Boosting—have demonstrated strong predictive performance and often outperform traditional approaches in discrimination metrics such as the area under the ROC curve (AUC). However, the adoption of machine learning in credit risk modeling remains debated due to concerns regarding probability calibration, temporal robustness, interpretability, and regulatory governance. This paper provides a comprehensive comparison of traditional statistical models and tree-based machine learning approaches in credit risk assessment. Rather than focusing exclusively on discriminatory performance, the analysis adopts a multidimensional evaluation framework incorporating calibration quality and temporal stability. Drawing on foundational theory and recent empirical evidence, the paper argues that model adequacy in credit risk is inherently context dependent. A three-pillar framework—discrimination, calibration, and temporal robustness—is proposed to guide academic research and practical model deployment. The findings suggest that superior ranking performance does not necessarily imply superior decision quality and that effective credit risk modeling requires balancing predictive flexibility with probabilistic reliability and governance stability.

Keywords

credit risk assessment, traditional statistical models, machine learning, logistic regression, tree-based models, model performance evaluation

1. Introduction

Credit risk modeling plays a central role in financial institutions by estimating the likelihood that a borrower will default on contractual obligations. These probability estimates directly influence loan approval decisions, risk-based pricing, expected loss calculation, portfolio risk management, and regulatory capital determination. Given the economic and systemic implications of these decisions, credit risk models must satisfy standards that extend beyond predictive accuracy.

For several decades, logistic regression has served as the benchmark modeling framework in credit scoring due to its probabilistic structure and operational compatibility with financial regulation [1]. Traditional risk management literature situates such models within broader institutional governance systems that emphasize validation, monitoring, and capital alignment [2].

With the expansion of digital financial data and computational capacity, machine learning techniques—particularly tree-based ensemble methods—have gained prominence in credit risk research. Empirical studies frequently report that these models achieve higher discrimination performance compared to logistic regression [3-5]. Despite these improvements, questions remain regarding calibration reliability, temporal robustness, and interpretability in regulated environments [6].

Despite extensive empirical comparisons between classical statistical models and machine learning algorithms, most of the extant literature relies on discriminatory performance metrics like AUC. The combination evaluation of calibration quality and temporal robustness has received very little attention, especially considering real-world deployment restrictions. This gap reduces the practical interpretability of model comparisons, highlighting the need for a more thorough evaluation methodology.

This paper synthesizes theoretical foundations and comparative empirical findings to clarify how traditional statistical models and machine learning approaches perform across dimensions that matter for real-world deployment rather than purely for predictive competition.

2. Practical Requirements of Credit Risk Models

Credit risk models operate in environments characterized by regulatory oversight, economic uncertainty, and high financial stakes. Regulatory guidelines emphasize that credit scoring systems must satisfy transparency, robustness, and governance requirements in addition to predictive strength [6].

First, models must generate reliable probability estimates. Default probabilities feed directly into expected loss (EL) and risk-weighted asset (RWA) calculations. Logistic regression's likelihood-based foundation ensures structural coherence between model specification and probabilistic interpretation [1].

Second, models must exhibit temporal stability. Borrower behavior and macroeconomic conditions evolve continuously, creating nonstationary data-generating processes. Effective credit risk management requires ongoing monitoring and recalibration to address performance deterioration over time [2].

Third, interpretability remains critical. Financial institutions must justify model decisions to regulators, auditors, and stakeholders. Models lacking explainable structure may encounter barriers to deployment regardless of predictive performance [6].

These practical constraints highlight the limitations of relying exclusively on predicted accuracy and encourage a more comprehensive evaluation approach. They explain why classical statistical models, particularly logistic regression, have remained prominent in credit risk modeling due to their probabilistic construction, interpretability, and regulatory compatibility, which directly fulfill these institutional expectations.

However, as credit data grows more complicated and multidimensional, it is vital to assess if existing models are still suitable. This prompts a closer examination of their strengths and weaknesses in modern credit risk contexts. From a methodological perspective, these requirements imply that credit risk modeling should be evaluated not only as a predictive task but also as a decision-support system embedded within institutional and regulatory frameworks. This perspective shifts the focus from model-centric performance to deployment-oriented adequacy.

3. Traditional Statistical Models in Credit Risk Assessment

3.1 Logistic Regression and Scorecard Frameworks

Logistic regression models default probability as a logistic transformation of a linear combination of explanatory variables. This structure provides clear probabilistic interpretation and supports regulatory transparency [1]. In applied settings, logistic regression is often implemented within scorecard frameworks that incorporate variable binning, monotonic constraints, and expert judgment. Such structured development processes enhance model stability and governance compatibility [2].

Logistic regression also benefits from structural simplicity, which can reduce sensitivity to noise and moderate distributional shifts. For example, logistic regression-based scorecards are frequently used in consumer lending scenarios like credit card approval or personal loan underwriting to estimate the likelihood of default using factors like income level, length of credit history, repayment patterns, and credit utilization ratios. To ensure economically significant linkages, these variables are usually modified using binning and monotonic restrictions; for example, larger debt-to-income ratios are consistently linked to increased default risk. Under regulatory examination, these structured modeling options improve interpretability and enable financial firms to give clear explanations for approval or rejection decisions.

Moreover, scorecard-based logistic regression models are frequently included into decision systems that convert estimated probabilities into loan limits, interest rates, or approval thresholds. Because these decisions have a direct impact on customer outcomes, the stability and calibration of anticipated probability are crucial. Even minor miscalibrations might result in a systematic over- or underestimation of risk, potentially resulting in financial losses or unfair lending practices.

Similarly, in small and medium enterprise (SME) lending, logistic regression models are commonly applied to structured financial indicators such as leverage ratios, liquidity measures, profitability metrics, and cash flow stability. In these situations, interpretability is especially crucial since credit officers and risk managers must examine model outputs alongside qualitative judgments such as management quality or industry perspective. Transparent models promote effective human-model interaction and auditability in regulatory reviews.

Furthermore, SME datasets are frequently smaller and noisier than large-scale consumer credit data, making simpler parametric models more resilient and less prone to overfitting. In such situations, the small predictive advantages provided by more complicated machine learning models may not be worth the sacrifice of interpretability and governance openness. As a result, classical logistic regression frameworks continue to play an important role in real-world credit risk modeling, despite the growing popularity of machine learning approaches.

3.2 Limitations in High-Dimensional and Nonlinear Settings

Despite its advantages, logistic regression imposes linear and additive assumptions that may limit its ability to capture complex nonlinear relationships and higher-order interactions among variables. In real-world credit risk data, borrower behavior is often influenced by intricate interactions—for example, the joint effect of income level, credit utilization, and macroeconomic conditions may not be well approximated by a purely linear specification.

As credit datasets expand in dimensionality and heterogeneity, traditional scorecard approaches may underfit unless extensive feature engineering is conducted. This process often requires manual variable transformation, interaction design, and domain expertise, making model development time-consuming and potentially subjective. Moreover, handcrafted features may fail to generalize across different datasets or changing economic environments, reducing model adaptability. Another limitation lies in the difficulty of capturing dynamic and evolving relationships in nonstationary data. As borrower characteristics and macroeconomic conditions shift over time, the fixed functional form of logistic regression may struggle to maintain predictive performance without frequent recalibration and redevelopment.

These limitations have led to more research into machine learning methods that can automatically find nonlinear interactions and complex patterns in large amounts of data. Such approaches offer greater flexibility and scalability, particularly in environments characterized by high dimensionality and rapidly changing data distributions.

4. Tree-Based Machine Learning Approaches

4.1 Ensemble Methods and Predictive Flexibility

Tree-based machine learning models partition the feature space through recursive splitting. Ensemble techniques such as Random Forest and Gradient Boosting aggregate multiple trees to improve generalization and reduce variance. Empirical comparisons consistently demonstrate improved discrimination performance for ensemble models relative to logistic regression [3-5]. In consumer and corporate credit datasets, boosting algorithms often achieve higher AUC values, particularly when nonlinear effects are pronounced.

In applied banking contexts, machine learning models have also been evaluated within Early Warning Systems (EWS). Evidence suggests that ensemble methods can enhance sensitivity to adverse client migration events and improve detection timeliness relative to traditional models [7]. These results highlight the value of machine learning in dynamic monitoring environments where early identification of deterioration is critical.

4.2 Probability Interpretation and Calibration Challenges

Despite strong discrimination performance, tree-based ensemble models are not inherently probabilistic. Their probability outputs are typically derived from averaged predictions rather than explicit likelihood-based estimation. Comparative studies frequently emphasize AUC while providing limited calibration diagnostics [3]. Uncalibrated machine learning models may produce biased or overly extreme probability estimates. Post-hoc calibration techniques can mitigate these issues but introduce additional modeling layers and potential sensitivity to temporal drift. In credit applications where probability magnitudes directly influence pricing and capital allocation, calibration quality constitutes a distinct and essential performance dimension.

5. Evidence from Comparative Studies

5.1 Discriminatory Performance

Importantly, these empirical findings should be interpreted within a broader decision-making context rather than as isolated performance comparisons. Across empirical studies, ensemble machine learning models often outperform logistic regression in discrimination metrics such as AUC and Gini coefficient [3-5]. However, the magnitude of these gains depends on evaluation methodology. Studies employing random train–test splits frequently report larger improvements than those implementing strict out-of-time validation frameworks [5]. This observation suggests that evaluation design significantly shapes conclusions about model superiority.

5.2 Calibration Performance

Calibration evidence presents a more nuanced picture. Logistic regression's parametric foundation supports coherent probability estimation [1]. Machine learning studies that incorporate calibration metrics often find that uncalibrated ensemble outputs require additional adjustment to achieve comparable reliability [3]. Because probability estimates serve as direct inputs into financial decision processes, systematic bias in predicted probabilities may distort expected loss projections even when ranking performance remains strong.

5.3 Temporal Stability and Drift

Temporal robustness has emerged as a central concern in recent research. Simulated macroeconomic stress scenarios indicate that boosting models may maintain strong discrimination under adverse conditions [4]. However, stability outcomes depend heavily on validation strategy and segmentation design [5]. In applied Early Warning Systems, machine learning models demonstrate improved detection of migration events but require monitoring infrastructure to sustain performance over time [7]. Temporal robustness therefore encompasses not only discrimination persistence but also probability stability and decision consistency under evolving economic conditions.

Taken together, these comparative findings indicate that differences in model performance are inherently multidimensional rather than reducible to a single metric. While machine learning models consistently demonstrate advantages in discrimination, particularly in complex and high-dimensional datasets, this superiority is often sensitive to evaluation design and does not uniformly translate into improvements in other performance dimensions. Calibration reliability and temporal stability exhibit greater variability across studies, suggesting that model performance must be interpreted within a broader evaluative context.

These observations are directly aligned with the central objective of this paper, which argues that credit risk model evaluation should extend beyond ranking-based metrics. In practical financial settings, predicted probabilities serve as direct inputs into downstream decision processes such as pricing, capital allocation, and risk provisioning. As a result, inaccuracies in probability estimation—especially systematic biases—can have material economic consequences even when discrimination performance remains strong. Furthermore, the presence of temporal drift highlights that models must maintain stability not only in predictive accuracy but also in probabilistic interpretation over time.

Importantly, the empirical evidence from comparative studies provides strong support for the three-pillar evaluation framework proposed in this paper—discrimination, calibration, and temporal robustness. Each pillar corresponds to a distinct dimension of model adequacy observed in the literature: discrimination captures the ability to rank borrowers by relative risk; calibration reflects the accuracy and reliability of predicted probabilities; and temporal robustness addresses the consistency of model performance under evolving data distributions and macroeconomic conditions.

From this perspective, comparative studies should not be interpreted as identifying a universally superior modeling approach. Rather, they reveal systematic trade-offs between predictive flexibility and probabilistic reliability. Consequently, effective credit risk modeling requires a multidimensional evaluation framework that integrates these complementary dimensions, aligning model selection with both statistical performance and real-world deployment requirements. This insight motivates the synthesis developed in the following section.

6. Discussion: A Three-Pillar Evaluation Framework

This study contributes to the existing literature by proposing a deployment-oriented three-pillar evaluation framework for credit risk modeling. Rather than viewing model comparison as a competition for predictive superiority, this framework emphasizes the need to align model performance with real-world decision requirements. The literature suggests that ensemble machine learning models frequently dominate discrimination metrics [3-5]. Logistic regression retains structural advantages in probabilistic coherence and governance compatibility [1, 6]. Stability outcomes depend on both algorithmic properties and institutional monitoring capacity [7].

Rather than framing model comparison as a binary contest, a deployment-oriented perspective recognizes that model adequacy depends on context. Machine learning models may be preferable in data-rich environments emphasizing ranking precision and early detection. Logistic regression remains competitive where probability reliability, interpretability, and regulatory transparency are paramount. Hybrid architectures that combine machine learning ranking with calibrated probability mapping represent a promising direction for integrating predictive flexibility with institutional reliability. From a theoretical standpoint, this paradigm adds to the literature by combining statistical modeling concepts with machine learning assessment procedures through a single, deployment-oriented lens.

However, there are some limits. First, dataset selection and validation design frequently influence empirical comparisons between research, which might have an impact on generalizability. Second, the relationship between model calibration and macroeconomic dynamics warrants more examination. Future study could look into hybrid modeling architectures that combine machine learning ranking skills with structurally grounded probabilistic models, as well as adaptive calibration techniques that improve temporal robustness during distributional transitions.

7. Conclusion

Credit risk modeling cannot be adequately evaluated using a single performance metric. While machine learning models frequently achieve superior discrimination performance, improvements in AUC do not necessarily translate into enhanced economic decision-making. Logistic regression continues to offer key advantages in probabilistic coherence, interpretability, and regulatory compatibility, which remain essential in institutional contexts.

This study contributes to the literature by proposing a three-pillar evaluation framework—discrimination, calibration, and temporal robustness—that captures the multidimensional nature of model adequacy. The analysis demonstrates that model selection should not be framed as a binary choice between traditional statistical methods and machine learning approaches, but rather as a context-dependent decision aligned with deployment requirements.

Future research should further explore hybrid modeling architectures that combine machine learning-based ranking with structurally grounded probability estimation, as well as adaptive calibration techniques that enhance robustness under distributional shifts. By emphasizing the integration of predictive performance with real-world constraints, this study provides a structured foundation for both academic research and practical implementation in credit risk modeling.

References

[1] Bolton, C. (2009). Logistic regression and its application in credit scoring. University of London.

[2] Brown, K., & Moles, P. (2008). Credit risk management. Edinburgh Business School, Heriot-Watt University.

[3] Xu, T. (2024). Comparative analysis of machine learning algorithms for consumer credit risk assessment. International Conference on Computational Intelligence and Applications (ICCIA).

[4] Vakrani, D. S., Padhye, P. S., Gupta, S. K., Roy, J. K., Nerlekar, V. S., & Parashar, N. (2026). Evaluating AI-driven credit scoring models versus traditional statistical techniques. Discover Artificial Intelligence, 6, 72. https://doi.org/10.1007/s44163-025-00772-1

[5] Oliveira, N. A. de, & Basso, L. F. C. (2025). Explaining corporate ratings transitions and defaults through machine learning. Algorithms, 18(10), 608. https://doi.org/10.3390/a18100608

[6] World Bank & International Committee on Credit Reporting. (2019). Credit scoring approaches guidelines. World Bank.

[7] Machado, M. R., Chen, D. T., & Osterrieder, J. R. (2025). An analytical approach to credit risk assessment using machine learning models. Decision Analytics Journal, 16, 100605. https://doi.org/10.1016/j.dajour.2025.100605

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access