Dean&Francis

ISSN 2959-6157

Research and Analysis of Explainable Machine Learning Methods for Credit Scoring and Loan Approval

Yiran Dai[1, *]

[1]College of Engineering and Applied Sciences, Stony Brook University, New York,11794, United States of America
*Corresponding author: daidaiqaq34@gmail.com

Abstract:

Financial institutions depend on credit scoring and loan approval systems to make essential decisions, as the accuracy of risk assessment directly affects asset performance, regulatory compliance, and fair treatment of consumers. Traditional statistical models remain widely used due to their simplicity and transparency, but their linear structure limits their ability to capture complex borrower behavior. Tree-based machine learning models, including Gradient Boosting Decision Trees, XGBoost, and Random Forest, offer stronger predictive performance; however, their explainability and transparency remain challenging in regulated credit settings. SHAP-based frameworks in Explainable Artificial Intelligence enable feature-level attribution that links model predictions to underlying financial characteristics. This survey reviews mainstream explainable tree-based credit scoring approaches, summarizes empirical findings, and discusses practical implementation challenges. It also outlines future directions related to regulatory requirements, fairness, drift adaptation, and scalable deployment. The results indicate that explainable tree-based scoring models can achieve strong performance while providing meaningful interpretability, although several unresolved challenges require further research.

Keywords: Credit scoring; Explainable machine learning; SHAP; XGBoost; Financial risk assessment

1. Introduction

Credit scoring is central to consumer and small business lending. A credit decision affects approval, credit limit, and interest rate for the applicant. At the portfolio level, decisions affect expected loss, capital planning, and risk appetite. In regulated environments, credit decisioning is not only a prediction task. It is also a compliance process. Institutions must satisfy model risk management expectations, provide audit-ready documentation, and communicate adverse action reasons to consumers in a consistent

way.

Logistic regression scorecards have been used for decades. They remain popular for practical reasons. The model form is simple. Coefficients can be interpreted as monotone effects under standard assumptions. Scorecards can be converted into points and reason codes that fit established governance workflows. This alignment with documentation and review is a key reason why scorecards are still dominant in many lenders [1]. Fintech and bank risk teams also value their stability under population changes, especially when feature engineering and binning are controlled.

The data environment has evolved. Lenders now use richer variables such as utilization dynamics, transaction patterns, and behavioral signals. Many relationships with default risk are non-linear. Risk may increase sharply after a threshold. Risk may also rise only when multiple conditions occur together. Linear scorecards can represent these effects only through extensive feature engineering. Feature engineering increases development time and can become fragile when the borrower population shifts. Macroeconomic changes can also alter the meaning of historical patterns. This makes stability a practical concern, not just a theoretical one [2].

Tree-based ensembles provide an alternative. Gradient boosting and Random Forest models can learn non-linear effects and interactions directly from the data. Many studies report that these models outperform logistic regression in discrimination metrics such as AUC and KS, especially when the feature space is complex or high-dimensional [2,3]. Credit scoring research has also proposed variants designed for risk modeling, including tree-enhanced boosting structures that improve ranking and calibration under credit data constraints [4]. In addition, operational studies increasingly evaluate not only default prediction but also profit scoring and decision optimization, where model outputs are used for pricing and acceptance control [5].

However, accuracy alone is not sufficient in credit risk. Tree ensembles are harder to justify because their logic is distributed across many trees. Regulators and internal validators must understand the model's behavior. Consumers must receive understandable adverse action reasons. Transparency and auditability have therefore become governance requirements rather than optional features [6]. Explainability is also closely linked to model monitoring. When data drift occurs, performance and explanation patterns can change. Institutions need tools to detect and document those changes.

Explainability alone does not guarantee fairness. A model can provide clear explanations and still produce biased outcomes if proxy variables correlate with protected characteristics. Responsible credit scoring therefore requires both interpretability and fairness evaluation, including subgroup performance analysis and mitigation procedures [7].

SHAP is one of the most widely used explanation methods for credit models. It is based on Shapley values from cooperative game theory and assigns feature contribution values for each prediction [8]. For tree-based models, TreeSHAP enables efficient computation and supports explanation at portfolio scale. SHAP values can be used for global model review as well as local decision explanations. In practice, however, deployment introduces challenges. Explanations may shift after retraining, reason codes may change, and governance policies must link explanation outputs to logging, monitoring, and operational workflows. These issues motivate a structured review of explainable tree-based credit scoring methods and their real-world constraints.

2. Main Techniques in Explainable Machine Learning Credit Scoring

Explainable credit scoring systems usually have two layers. The predictive layer outputs a probability of default or a risk score. The explanation layer produces reason signals for audit, review, and consumer communication. Tree-based ensembles are common in credit scoring because they capture complex patterns. Post hoc explanation methods are then used to make their behavior interpretable.

2.1 Gradient Boosted Decision Trees

Gradient Boosting Decision Trees build an ensemble sequentially. Each new tree learns to correct errors made by the previous trees. This is useful for credit scoring because defaults are rare and borderline cases near decision thresholds are common. Boosting can focus model capacity on difficult cases and improve ranking performance.

A tree-enhanced gradient boosting approach has been proposed for credit scoring and reported improvements over logistic regression baselines on credit datasets that resemble production settings [4]. The business value of small ranking improvements can be large. Better ranking supports better cutoff selection. It can also improve pricing and acceptance control. However, increased complexity makes the model harder to interpret directly. This is a key reason why explainability tools are paired with boosting models in regulated credit.

2.2 XGBoost

XGBoost is a widely used boosting implementation. It includes regularization, efficient training, and support for parallel processing. These properties make it attractive in industry where large datasets and frequent retraining

ISSN 2959-6157

cycles are common. In credit risk studies, XGBoost often improves discrimination over linear scorecards while remaining feasible for production deployment [2].

A Journal of Financial Data Science study presents explainable machine learning workflows for consumer credit risk and shows how multiple stakeholders can use explanations during development and review [2]. In practice, XGBoost is often combined with TreeSHAP for feature attribution. Global SHAP summaries support feature sanity checks and model validation. Local SHAP values can support case-level review and adverse action reasoning. Yet institutions must manage explanation stability. Even when AUC remains similar, feature attributions can change across retraining cycles. This can affect reason code consistency and documentation. Version control and explanation logging are therefore operational requirements, not optional improvements.

2.3 Random Forest Models

Random Forest trains many trees independently and aggregates their outputs. This reduces variance and can improve robustness under noisy data conditions. Credit datasets often contain outliers, missing values, and mixed populations. Random Forest can be a strong baseline in these settings, and many comparative studies include it as a reference model [3].

Random Forest may underperform boosting methods in rare default detection when class imbalance is severe. Calibration and threshold selection can also affect usefulness for decisioning. Like other ensembles, Random Forest is not inherently interpretable at the model level. Post hoc methods are commonly used. SHAP can be applied to Random Forest to produce local and global explanations, but explanation cost can be higher than for simpler models.

2.4 Hybrid Interpretation-Enhanced Structures

A central operational challenge is translating model explanations into reason codes. Institutions must provide concise consumer-facing reasons and maintain stable categories for governance. Hybrid approaches try to bridge this gap.

One approach proposes using Shapley values to construct interpretable credit scorecards that resemble traditional workflows while leveraging machine learning signal [4]. This framework aims to align explanation outputs with familiar scorecard artifacts, which can reduce friction in governance and compliance. Another direction constrains model structure to improve interpretability. For example, inherently interpretable classification tree approaches with more flexible split structures can improve interpretability while retaining non-linear decision boundaries [9]. Additive boosting variants have also been proposed to preserve a closer-to-additive form that is easier to justify and monitor [10].

Other methods enhance boosting using auxiliary learners or alternative training strategies. An Extreme Learning Machine enhanced gradient boosting method represents one example of how researchers attempt to improve performance while keeping models practical [11]. These methods may offer gains on some datasets, but they still require explanation layers and careful validation.

Hybrid explainability frameworks increasingly include fairness components. A Scientific Reports study proposes an explainable hybrid method with transparency and fairness considerations, highlighting that explainability should be paired with bias assessment rather than treated as a standalone solution [12]. At a broader level, fintech risk management literature emphasizes that explainability should be integrated into risk governance, not bolted on after modeling [13]. Earlier applied work on explainability in finance also shows how explanations can be used for default risk analysis while emphasizing limitations and governance needs [14]. General XAI surveys further classify explanation methods, their goals, and common failure modes, which helps contextualize SHAP and hybrid approaches in deployment settings [15].

In summary, modern explainable credit scoring often combines high-performing tree ensembles with SHAP-based attribution and governance-oriented translation layers. The technical toolbox is mature, but deployment quality depends on monitoring, stability management, and operational mapping into reason codes.

3. Empirical Findings and Result Observations

Across recent studies, tree-based ensembles often show stronger discrimination than logistic regression scorecards. The size of improvement depends on the dataset, feature set, and evaluation protocol. Many studies report gains of a few AUC points. In large portfolios, even small ranking improvements can translate into meaningful business value. Better ranking supports improved cutoff selection and pricing strategies while controlling approval rates.

A representative example is the Journal of Financial Data Science study on explainable models for consumer credit risk. It demonstrates machine learning models, explanation workflows, and stakeholder use cases, emphasizing that explanations support both development and governance [2]. Studies focusing on imbalanced credit datasets also show that modeling choices can improve minority class detection and maintain interpretability [3]. A European Journal of Operational Research paper analyzes interpretable machine learning for imbalanced credit scoring and highlights evaluation choices that matter in practice

[3]. These findings support the view that explainable machine learning can be both accurate and useful, but only if training and validation are aligned with credit decision objectives.

Class imbalance is a core issue. Defaults are rare. If the training process is not handled carefully, a model may focus on majority-class patterns and miss rare default drivers. This affects both performance and explanations. When a model learns meaningful default drivers, SHAP attributions tend to highlight those drivers more consistently. When a model fits spurious correlations, explanations may look plausible but fail to generalize. This is why empirical evaluation should consider more than AUC. Credit risk studies often include calibration, stability checks, and subgroup analysis.

SHAP explanations often align with domain intuition. Common global drivers include utilization, delinquency history, payment behavior, and income stability. Global SHAP summaries can help validators confirm that the model uses plausible signals. Local SHAP values can support case reviews and help teams prepare adverse action explanations. A Shapley-based scorecard framework provides a concrete example of how to bridge SHAP attributions with operational artifacts and reason categories [4]. In practice, institutions may group features into reason families and then map the highest-impact groups into standardized reason codes.

Explanation stability over time is another key observation. Credit models are retrained and borrower populations shift. Economic regimes can change. During stress periods, variables linked to payment burden or leverage may become more influential. Even if overall AUC remains stable, explanation patterns may shift. This can reflect real changes in risk mechanisms. It can also reflect instability from data refreshes or tuning changes. Governance research emphasizes that auditability and explainability should be evaluated across the model lifecycle, including monitoring and documentation practices [6].

Fairness is also central in empirical deployment. Explainability does not prevent bias. Proxy variables can lead to disparate outcomes. Responsible machine learning practices emphasize subgroup performance evaluation, monitoring, and mitigation decisions [7]. Hybrid work that explicitly combines transparency with fairness objectives provides evidence that fairness must be treated as a governance requirement and tested continuously [12]. Explanation patterns can also differ across subgroups, and such differences can provide diagnostic signals for potential proxy effects.

Beyond default prediction, empirical work also considers decision objectives such as profit scoring. Evidence from credit unions and other lenders compares algorithms such as multilayer perceptron, XGBoost, and TabNet. These studies emphasize that model selection should depend on business objectives, data constraints, and explainability needs, rather than accuracy alone [16]. More broadly, fintech research highlights that explanation tools add value when integrated with risk management practices, dashboards, and human review processes [13]. Applied financial explainability work also shows the role of explanations in understanding default risk while warning against over-trusting post hoc explanations [14]. General XAI frameworks summarize common explanation goals and risks, which helps interpret empirical results and avoid misuse [15]. Finally, explanation accuracy itself can be evaluated. A Banco de España working paper proposes methods to assess whether explanations correctly identify true drivers in credit decision contexts, which supports supervisory-style assessment [17].

Overall, empirical evidence supports explainable tree ensembles as practical tools for credit scoring. The strongest results occur when institutions align training with imbalance constraints, monitor stability, and validate explanation quality alongside predictive performance.

4. Discussion

Explainable tree-based credit scoring faces several deployment barriers. The first barrier is computational cost. TreeSHAP is efficient, but large-scale explanation generation can still be expensive for big portfolios. Institutions often restrict full explanations to adverse decisions, audit samples, or targeted investigations. They may compute global explanation summaries on a scheduled cadence and generate local explanations on demand.

The second barrier is stability and documentation. Models are retrained, data pipelines change, and explanation outputs can shift. Changes in attribution rankings can lead to changes in reason codes and communication templates. This creates audit risk and customer communication risk. Robust version control, reproducible training pipelines, and explanation logging are necessary to support governance [6]. Institutions should store model versions, feature definitions, and explanation outputs together so that decisions can be reconstructed during reviews.

The third barrier is fairness. Explainability does not remove bias by itself. Proxy variables can produce disparate impacts even when explanations are clear. Responsible machine learning guidance emphasizes that lenders must test subgroup performance, monitor fairness metrics over time, and document mitigation decisions and tradeoffs [7]. Hybrid approaches that include fairness considerations reinforce that fairness should be treated as part of the system design, not as an afterthought [12]. Institutions should also analyze subgroup differences in explanation patterns, since shifts in explanations can indicate proxy effects or

ISSN 2959-6157

drift.

The fourth barrier is communication. Different stakeholders require different explanation forms. Consumers need short, understandable reasons. Regulators and validators require detailed technical documentation. Developers need diagnostic explanations for debugging and monitoring. This requires translation from numeric attributions into stable reason categories. It also requires clear feature grouping rules and validation of the mapping procedure. Explanation accuracy frameworks can help validate whether the explanation approach identifies true drivers under controlled conditions, which strengthens audit defensibility [17]. Treating explainability as a lifecycle system, rather than a single metric, is consistent with broader fintech and XAI literature [13,15].

5. Conclusion

Tree-based models such as GBDT, XGBoost, and Random Forest frequently improve credit scoring performance by capturing non-linear effects and feature interactions. SHAP and TreeSHAP provide practical explanation tools that support global model review and local decision explanations. Empirical studies show that these methods can be deployed in credit scoring workflows and can support auditing, monitoring, and stakeholder communication.

However, deployment requires more than selecting a model and computing SHAP values. Institutions must manage explanation stability across retraining cycles, map attributions into consistent reason codes, and maintain audit-ready documentation. Fairness risks must be evaluated continuously, since explainability does not eliminate proxy effects or disparate impact. Monitoring should include both performance drift and explanation drift. Future work should improve drift-aware and stability-aware explanation methods, integrate fairness constraints into training and governance, and develop stronger standards for validating explanation accuracy in credit decision settings. With these safeguards, explainable tree ensembles can support responsible lending decisions while maintaining strong predictive performance.

References

[1] Demajo L, Vella V, Dingli A. Explainable AI for interpretable credit scoring. arXiv:2012.03749, 2020.

[2] Davis R, Lo AW, Mishra S, Nourian A, Singh M, Wu N, Zhang R. Explainable Machine Learning Models of Consumer Credit Risk. The Journal of Financial Data Science, 2023, 5(4): 9–39. DOI: 10.3905/jfds.2023.1.141.

[3] Chen Y, Calabrese R, Martin-Barragán B. Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research, 2024.

[4] Liu W, Fan H, Xia M. Credit scoring based on tree-enhanced gradient boosting decision trees. Expert Systems with Applications, 2022, 189: 116034. DOI: 10.1016/j.eswa.2021.116034.

[5] Asencios R, Asencios C, Ramos E. Profit scoring for credit unions using the multilayer perceptron, XGBoost and TabNet algorithms: Evidence from Peru. Expert Systems with Applications, 2023, 213: 119201. DOI: 10.1016/j.eswa.2022.119201.

[6] Bücker M, Szepannek G, Gosiewska A, Biecek P. Transparency, auditability, and explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 2022, 73(1): 70–90. DOI: 10.1080/01605682.2021.1922098.

[7] Valdrighi G, Ribeiro AM, Pereira JSB, et al. Best practices for responsible machine learning in credit scoring. Neural Computing & Applications, 2025, 37(25): 20781–20821. DOI: 10.1007/s00521-025-11520-y.

[8] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, 30. DOI: 10.48550/arXiv.1705.07874.

[9] Tu J, Wu Z. Inherently interpretable machine learning for credit scoring: Optimal classification tree with hyperplane splits. European Journal of Operational Research, 2025. DOI: 10.1016/j.ejor.2024.10.046.

[10] Zou Y, Xia M, Lan X. Interpretable credit scoring based on an additive extreme gradient boosting. Chaos, Solitons & Fractals, 2025, 194: 116216. DOI: 10.1016/j.chaos.2025.116216.

[11] Zou Y, Gao C. Extreme Learning Machine Enhanced Gradient Boosting for Credit Scoring. Algorithms, 2022, 15(5): 149. DOI: 10.3390/a15050149.

[12] Nwafor CN, Nwafor O, Brahma S. Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach. Scientific Reports, 2024.

[13] Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable AI in fintech risk management. Frontiers in Artificial Intelligence, 2020, 3: 26. DOI: 10.3389/frai.2020.00026.

[14] Bracke P, Datta A, Jung C, Sen S. Machine learning explainability in finance: an application to default risk. Bank of England Staff Working Paper, 2019, No. 816.

[15] Arrieta AB, Del Ser J, et al. Explainable Artificial Intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82–115.

[16] Hlongwane R, Ramabao K, Mongwe W. A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. PLOS ONE, 2024, 19(8): e0308718.

[17] Alonso A, Carbó JM. Accuracy of explanations of machine learning models for credit decisions. Banco de España Working Papers, 2022, No. 2222.