International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct-2026, ISSN (Online): 2395-566X

Explainable AI For Credit Risk Assessment: A Review Of Machine Learning Models, Post-Hoc Explanation Techniques And The Accuracy–Interpretability Trade-Off

Irakshetti Pranav Amol; Prof. C. A. Kolse; Prof. M. C. Kolse; Prof. S. K. Said
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India;

Abstract—Credit risk assessment is an important process in the banking and financial industry, as it helps financial institutions evaluate the possibility of borrowers failing to repay their loans. With the advancement of Artificial Intelligence (AI) and Machine Learning (ML), traditional credit assessment methods are increasingly being supported by intelligent prediction models. However, many advanced AI models operate as black-box systems, making it difficult to understand how they arrive at particular credit decisions. Explainable Artificial Intelligence (XAI) addresses this limitation by providing transparency and interpretability in AI-based credit risk assessment. This paper presents a review of existing XAI techniques and their applications in predicting credit risk and supporting financial decision-making. Various machine learning approaches, including Decision Trees, Random Forest, Logistic Regression, and Gradient Boosting, are examined along with explainability techniques such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations), and feature importance analysis. The review discusses how these methods help identify important factors influencing credit scores, loan approvals, and default predictions. It also highlights the benefits of explainable models, including improved transparency, customer trust, accountability, and regulatory compliance. Furthermore, major challenges such as data privacy, algorithmic bias, model complexity, and the balance between prediction accuracy and interpretability are explored. The paper also discusses future research opportunities in developing fair, reliable, and transparent AI-based financial systems. Overall, this review emphasizes the importance of Explainable AI in making credit risk assessment more understandable, trustworthy, and responsible while supporting better financial decisions.

Index Terms—Explainable artificial intelligence, credit risk assessment, machine learning, SHAP, LIME, credit scoring, loan default prediction, model interpretability, financial decision-making

I. INTRODUCTION

When a loan application is rejected, the applicant is entitled to know why. A traditional scorecard can answer this directly: it is a weighted sum of a few variables, and the officer can point to the ones that lowered the score. A gradient boosting model built on hundreds of features cannot answer in the same way, even though it may predict default more accurately.

This tension sits at the centre of modern credit risk assessment. Hand and Henley [1] reviewed statistical classification for consumer credit scoring when logistic regression and discriminant analysis dominated, and Thomas [2] surveyed credit and behavioural scoring as a forecasting problem. Baesens *et al.* [3] benchmarked classifiers on real credit datasets and found that although advanced methods performed well, simple models remained competitive. Two decades later, Dastile *et al.* [4] systematically reviewed statistical and machine learning models for credit scoring and reported that ensembles consistently lead on accuracy, while noting that most studies ignore interpretability entirely.

Explainable Artificial Intelligence (XAI) attempts to recover the lost transparency without giving up accuracy. This paper reviews machine learning models used for credit risk, the explanation techniques applied to them, what those explanations can and cannot support, and how institutions can combine them into a deployable, auditable workflow.

II. BACKGROUND AND RELATED WORK

2.1 Credit Risk Decisions

Credit risk assessment answers a probability question: given an applicant's financial history, income, existing obligations and behaviour, what is the probability of default within a defined horizon? The model output is converted into a score, a decision (approve, reject, refer) and often a price (interest rate or credit limit). Each of these downstream uses places a different demand on explanation.

2.2 Interpretability And Explainability

An *interpretable* model can be understood directly from its structure, such as a logistic regression scorecard or a shallow decision tree. An *explainable* system uses a separate technique to describe the behaviour of a model that is not itself understandable. Doshi-Velez and Kim [5] argued for a rigorous definition of interpretability tied to the task, and Guidotti *et al.* [6] surveyed methods for explaining black-box models. Rudin [7] went further, arguing that for high-stakes

© 2026 IJSRET

decisions such as lending, an inherently interpretable model should be preferred to an explanation of a black box.

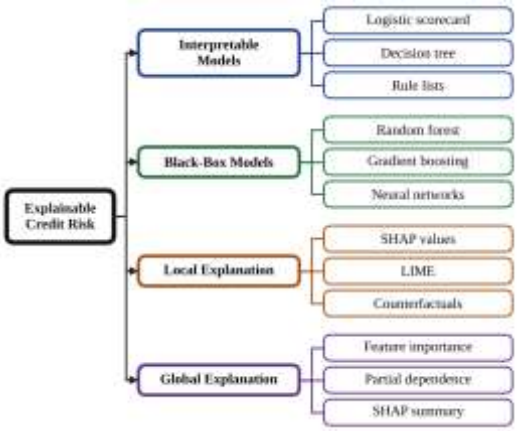


Fig. 1 classifies the approaches reviewed in this paper.

Fig. 1. Approaches to transparency in credit risk models.

2.3 Review Of Related Studies

From scorecards to ensembles. Altman [8] introduced discriminant analysis for bankruptcy prediction, establishing quantitative credit assessment. Logistic regression became the industry standard because its coefficients translate directly into scorecard points. Breiman [9] introduced random forests, and Friedman [10] formulated gradient boosting, later made scalable by XGBoost [11] and LightGBM [12]. Dastile *et al.* [4] found ensembles to be the most accurate family across credit datasets, and Gunnarsson *et al.* [13] compared deep learning with gradient boosting on credit scoring data and concluded that deep networks did not outperform boosted trees on typical tabular credit data, which matters because the added opacity brings no accuracy benefit.

Explanation techniques. Ribeiro *et al.* [14] proposed LIME, which fits a simple local model around one prediction to approximate the black box in that neighbourhood. Lundberg and Lee [15] proposed SHAP, which assigns each feature a contribution based on Shapley values from cooperative game theory and satisfies consistency properties that ad-hoc importance measures do not. Friedman [10] also introduced partial dependence, which shows the average effect of a feature across the population. Molnar [16] provides a consolidated treatment of these methods.

Applied to credit risk. Bracke *et al.* [17] applied Shapley-based explanations to a mortgage default model at the Bank of England and showed that the method can produce explanations usable by model reviewers and supervisors. The FICO Explainable Machine Learning Challenge [18] released the HELOC dataset specifically to test whether accurate credit models can also be explainable; Chen *et al.* [19] responded with a globally consistent interpretable model that matched black-box accuracy on that data, evidence that the trade-off is often smaller than assumed.

Ref.	Author(s) / Year	Technique / Approach	Key Contribution	Limitation / Gap
[8]	Altman, 1968	Discriminant bankruptcy model	Founded quantitative credit scoring	Corporate, few ratios
[1]	Hand and Henley, 1997	Review of statistical scoring	Mapped classical methods	Pre-ensemble era
[3]	Baesens et al., 2003	Benchmark of classifiers	Simple models competitive	Small datasets
[14]	Ribeiro et al., 2016	LIME local surrogate	Model-agnostic explanation	Unstable locally
[15]	Lundberg and Lee, 2017	SHAP attributions	Consistent additive credit	Compute cost
[19]	Chen et al., 2018	Interpretable credit model	Matched black-box accuracy	One dataset
[17]	Bracke et al., 2019	Shapley on default risk	Supervisor-usable explanations	Single portfolio
[20]	Slack et al., 2020	Attacks on LIME and SHAP	Explanations can be faked	Adversarial setting
[4]	Dastile et al., 2020	Systematic credit scoring review	Ensembles lead on accuracy	Interpretability ignored
[13]	Gunnarsson et al., 2021	Deep learning vs. boosting	No gain from deep nets	Tabular data only
[24]	Fuster et al., 2022	ML effects on credit markets	Unequal distribution of gains	US mortgages

TABLE I Summary of Literature Survey

Limits of post-hoc explanation. Slack *et al.* [20] constructed models that behave in a biased way yet produce innocuous LIME and SHAP explanations, demonstrating that post-hoc explanations can be deliberately manipulated. Alvarez-Melis and Jaakkola [21] showed that explanations can change substantially under small input perturbations, and Kumar *et al.* [22] analysed conceptual problems with interpreting Shapley values as feature importance. Barocas *et al.* [23]

© 2026 IJSRET

examined the assumptions hidden in counterfactual explanations and principal reasons given to applicants.

Fairness and regulation. Fuster *et al.* [24] studied the effect of machine learning on US mortgage markets and found that more flexible models improved predictions overall but distributed gains unevenly across racial groups, a central concern for credit. In the United States, Regulation B under the Equal Credit Opportunity Act [25] requires specific principal reasons for adverse action, effectively mandating explanation. The GDPR [26] and the EU AI Act [27] impose transparency and risk-management obligations, supervisory guidance such as SR 11-7 [28] governs model risk management, and in India the Reserve Bank's digital lending guidelines [29] require transparency in automated loan decisions. Table I summarizes the reviewed works.

III. RESEARCH GAP AND OBJECTIVES

3.1 Identified Research Gaps

- Most credit scoring studies report accuracy only and omit any interpretability assessment.
- LIME and SHAP explanations can be unstable or adversarially manipulated, yet are used as compliance evidence.
- There is no standard metric for the quality of a credit decision explanation.
- Explanations produced for model developers are rarely usable by loan applicants.
- Accuracy and fairness improvements are reported in aggregate, hiding unequal effects across groups.
- Indian credit datasets with thin-file and informal-income borrowers are scarce in published research.

3.2 Objectives Of The Review

- **O1:** To review machine learning models used for credit risk assessment and compare their predictive performance and transparency.
- **O2:** To examine post-hoc explanation techniques, particularly SHAP, LIME and feature importance, in the credit context.
- **O3:** To assess the reliability and known failure modes of these explanation techniques.
- **O4:** To relate explainability to fairness, customer trust and regulatory requirements.
- **O5:** To propose a layered architecture for explainable credit risk assessment.
- **O6:** To identify challenges and future research directions.

IV. DESIGN OF THE PROPOSED SYSTEM

Fig. 2 presents the proposed five-layer architecture, in which explanation is generated and validated as part of the decision, not added afterwards.



Fig. 2. Proposed five-layer architecture for explainable credit risk assessment.

4.1 Data Layer

Application data, credit bureau history, repayment behaviour, banking transactions and declared income are collected with consent. Protected attributes are recorded separately for fairness testing but excluded from the model features, and data quality and lineage are tracked as required by risk data aggregation principles [30].

4.2 Modelling Layer

Two models are trained in parallel: an interpretable benchmark (regularized logistic regression scorecard or a shallow tree) and a challenger ensemble (random forest or gradient boosting). The challenger is adopted only if its performance gain over the benchmark is material on held-out and out-of-time data.

4.3 Explanation Layer

For each decision, SHAP values give the contribution of every feature; the largest negative contributions become the principal reasons. Partial dependence and global SHAP summaries describe overall model behaviour for reviewers. A counterfactual module identifies the smallest realistic change that would alter the outcome, restricted to features the applicant can act on.

4.4 Validation And Fairness Layer

Explanation stability is tested by re-computing explanations under small perturbations and across retrained models; agreement between the benchmark and challenger explanations is checked; approval rates and error rates are compared across groups; drift in features and explanations is monitored.

4.5 Decision And Disclosure Layer

The officer sees the score, the principal reasons, the counterfactual and the confidence. The applicant receives a plain-language reason statement and guidance. Every decision, model version and explanation is logged for audit and appeal.

4.6 Methodology

© 2026 IJSRET

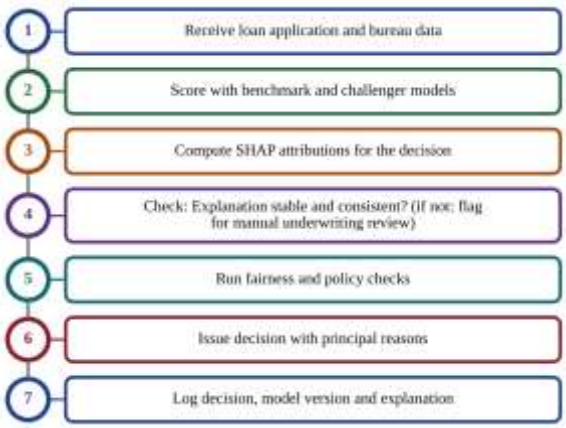


Fig. 3 describes the working of the system from application to disclosure.

Fig. 3. Working of the explainable credit risk assessment process.

4.7 Prediction

The probability of default for applicant x is

$$PD = P(default \mid x) = f(x) \quad (1)$$

and the decision follows a threshold set from the institution's loss and approval policy.

4.8 Attribution

SHAP decomposes a single prediction into an additive sum of feature contributions,

$$f(x) = \varphi_0 + \Sigma_j \varphi_j \quad (2)$$

where φ_0 is the average model output and φ_j the contribution of feature j. The features with the largest negative φ_j are reported as the principal reasons for rejection. LIME instead fits a local surrogate g by minimizing

$$L(f, g, \pi_x) + \Omega(g) \quad (3)$$

where π_x weights samples near x and Ω penalizes complexity of g.

4.9 Evaluation

Discrimination is measured with AUC and the Gini coefficient,

$$Gini = 2 \cdot AUC - 1 \quad (4)$$

Explanation stability is measured by the agreement of the top-k reasons under perturbation,

$$Stability = |R_k(x) \cap R_k(x')| / k \quad (5)$$

and fairness by comparing approval and default rates across groups. A model is approved for deployment only when accuracy, stability and fairness thresholds are all met.

V. DISCUSSION AND COMPARISON

Table II compares the model families on predictive performance, native transparency, explanation requirement and typical use in lending. Table III then summarizes the challenges.

Model	Accuracy	Native Transparency	Explanation Needed	Typical Use
Logistic scorecard	Moderate	High	None	Regulatory benchmark
Decision tree	Moderate	High (if shallow)	None	Policy rules
Random forest	High	Low	SHAP / importance	Challenger model
Gradient boosting	Highest	Low	SHAP / LIME	Primary scoring
Neural network	High	Very low	SHAP / LIME	Rarely justified

TABLE II Comparison of Models for Credit Risk Assessment

VI. CHALLENGES AND FUTURE DIRECTIONS

Explanations are only useful if they are stable, honest and understandable to the person affected. Current techniques satisfy these conditions imperfectly, and the evidence from Slack *et al.* [20] shows that they can be subverted by a model owner who wishes to hide behaviour. Future work includes standardized benchmarks for explanation quality in credit, interpretable-by-design scoring models in the direction of Chen *et al.* [19], fairness-constrained training, privacy-preserving learning across institutions, and explanation interfaces evaluated with real applicants rather than with developers.

TABLE III Challenges and Mitigation

© 2026 IJSRET

Challenge	Mitigation
Accuracy–interpretability trade-off	Interpretable benchmark; adopt challenger only on material gain
Unstable or manipulated explanations	Perturbation and cross-model stability tests
Algorithmic bias	Group metrics; fairness-constrained training; bias audits
Data privacy	Consent, minimization, DPDP Act compliance
Model complexity and drift	Version control; periodic revalidation
Explanations unusable by applicants	Plain-language reasons; actionable counterfactuals

VII. CONCLUSIONS

Explainable AI makes credit risk assessment more understandable, trustworthy and accountable. The reviewed literature shows that ensemble models such as random forests and gradient boosting lead on predictive accuracy for credit data, that SHAP and LIME can expose the factors behind individual decisions, and that regulators increasingly require such explanation. It also shows real limits: post-hoc explanations can be unstable or manipulated, and accuracy gains from complex models may be distributed unequally across borrowers. The practical conclusion is to keep an interpretable benchmark alongside any complex model, to validate explanations as rigorously as predictions, and to treat explainability as a requirement of responsible lending rather than as a reporting formality.

7.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. D. J. Hand and W. E. Henley, “Statistical classification methods in consumer credit scoring: A review,” J. Roy. Statist. Soc. A, vol. 160, no. 3, pp. 523–541, 1997.
2. L. C. Thomas, “A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers,” Int. J. Forecasting, vol. 16, no. 2, pp. 149–172, 2000.
3. B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, “Benchmarking state-of-the-art classification algorithms for credit scoring,” J. Oper. Res. Soc., vol. 54, no. 6, pp. 627–635, 2003.
4. X. Dastile, T. Celik, and M. Potsane, “Statistical and machine learning models in credit scoring: A systematic literature survey,” Appl. Soft Comput., vol. 91, Art. no. 106263, 2020.
5. F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv:1702.08608, 2017.
6. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” ACM Comput. Surv., vol. 51, no. 5, Art. no. 93, 2018.
7. C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019.
8. E. I. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” J. Finance, vol. 23, no. 4, pp. 589–609, 1968.
9. L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
10. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.
11. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.
12. G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” in Proc. NeurIPS, 2017, pp. 3146–3154.
13. B. R. Gunnarsson, S. vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, “Deep learning for credit scoring: Do or don’t?,” Eur. J. Oper. Res., vol. 295, no. 1, pp. 292–305, 2021.
14. M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?”: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.
15. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. NeurIPS, vol. 30, 2017.
16. C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
17. P. Bracke, A. Datta, C. Jung, and S. Sen, “Machine learning explainability in finance: An application to default risk analysis,” Bank of England Staff Working Paper No. 816, 2019.
18. FICO, “Explainable Machine Learning Challenge (HELOC dataset),” 2018. [Online]. Available:

© 2026 IJSRET

https://community.fico.com/s/explainable-machine-learning-challenge

19. C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "An interpretable model with globally consistent explanations for credit risk," arXiv:1811.12615, 2018.
20. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proc. AAAI/ACM Conf. AI, Ethics, Society (AIES), 2020, pp. 180–186.
21. D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," arXiv:1806.08049, 2018.
22. I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in Proc. 37th ICML, 2020, pp. 5491–5500.
23. S. Barocas, A. D. Selbst, and M. Raghavan, "The hidden assumptions behind counterfactual explanations and principal reasons," in Proc. Conf. Fairness, Accountability, Transparency (FAT)*, 2020, pp. 80–89.
24. A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably unequal? The effects of machine learning on credit markets," J. Finance, vol. 77, no. 1, pp. 5–47, 2022.
25. Consumer Financial Protection Bureau, "Equal Credit Opportunity Act (Regulation B), 12 CFR Part 1002—Adverse action notification requirements," Washington, DC, USA.
26. European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Off. J. Eur. Union, L 119, 2016.
27. European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Off. J. Eur. Union, 2024.
28. Board of Governors of the Federal Reserve System, "Supervisory guidance on model risk management," SR Letter 11-7, 2011.
29. Reserve Bank of India, "Guidelines on digital lending," RBI/2022-23/111, Mumbai, India, Sep. 2022.
30. Basel Committee on Banking Supervision, "Principles for effective risk data aggregation and risk reporting," BCBS 239, Bank for International Settlements, Basel, Switzerland, 2013.

© 2026 IJSRET