The American Journal of Engineering and Technology
ISSN 2689-0984 Volume 07 - 2025

Interpretable AI in Credit Scoring: A Comparative Survey of SHAP, LIME, and Hybrid Approaches

[1]Sai Prashanth Pathi, [2]Jahnavi Swetha Pothineni
[1,2]Independent Researcher, USA

Received: 13th Sep 2025 | Received Revised Version: 21th Oct 2025 | Accepted: 21th Nov 2025 | Published: 29th Nov 2025

Volume 07 Issue 11 2025 | Crossref DOI: 10.37547/tajet/v7i11-304

Abstract

Explainable AI (XAI) is critical in domains like credit scoring where model decisions must be transparent and accountable. This survey paper compares three local explanation techniques—SHAP, LIME, and ensemble Hybrid approach that integrates both. We evaluate these methods on consistency, variability, and suitability for regulatory environments. Emphasis is placed on use in credit risk modeling, with insights drawn from both literature and practical evaluation.

Keywords: Explainable AI (XAI), SHAP, LIME, Local interpretability, Hybrid model explanations, Credit Risk Modeling.

Cite This Article: Pathi, S. P., & Pothineni, J. S. (2025). Interpretable AI in credit scoring: A comparative survey of SHAP, LIME, and hybrid approaches. The American Journal of Engineering and Technology, 7(11), 151–155. https://doi.org/10.37547/tajet/v7i11-304.

1. Introduction

Although black-box models dominate machine learning applications in finance, their lack of interpretability limits their suitability for deployment. XAI methods help address this challenge. This paper surveys two prominent local explanation methods SHAP and LIME, and a hybrid technique combining them. Our focus is on their comparative performance in terms of explanation quality, stability, and repeatability in high-stakes environments.

2. Background And Motivation

Modern financial institutions increasingly adopt black-box machine learning models like Random Forests, Gradient Boosting Machines and Neural Nets for tasks such as credit scoring. While these models offer high predictive accuracy, they lack transparency, posing challenges in regulated and ethically sensitive environments. Regulatory frameworks like GDPR and Fair Lending laws require decisions to be explainable, particularly when they impact individuals' financial access. Interpretability is therefore essential not only for compliance but also to foster trust and detect potential biases. To address this, local explanation methods are widely used to interpret individual model predictions. Two prominent techniques are:

- **LIME (Local Interpretable Model-agnostic Explanations):** Constructs a local surrogate model by perturbing input features around a data instance and fitting a simple model (e.g., linear regression) to approximate the complex decision boundary. While efficient and intuitive, LIME suffers from sensitivity to

randomness, often resulting in inconsistent explanations.

- **SHAP (SHapley Additive exPlanations):** Computes feature contributions based on Shapley values from co- operative game theory, ensuring consistency and globalinterpretability. However, SHAP can be computationally expensive and may make simplifying assumptions, such as feature independence.

Given their complementary properties LIME's local adaptability and SHAP's theoretical rigor, hybrid methods that combine both have emerged as promising solutions. This paper explores and evaluates such a hybrid approach in addition to the individual techniques.

3. Literature Review

Ribeiro et al. [1] introduced LIME as a local surrogate-based method that uses random perturbations to generate explanations, though it suffers from explanation variability. Lundberg et al. [2] proposed SHAP, using Shapley values from cooperative game theory to ensure consistent feature attribution. Recent work by Slack et al. [3] and Alvarez et al. [4] demonstrated adversarial vulnerabilities in both LIME and SHAP. Krishna et al. [5] studied disagreement among local methods across real-world applications. Carta et al. [6] and Vilone et al. [7] conducted broader reviews highlighting the strengths and deployment challenges of XAI in finance. Bhatt et al. [8] discussed practical hurdles in operationalizing XAI. Counterfactual approaches by Mothilal et al. [9] were also proposed to improve fairness in ML models. Our work synthesizes this literature and adds empirical evaluation of a hybrid SHAP-LIME framework which is an ensemble of both that can be used for credit scoring scenarios.

4. Methodology

To comprehensively evaluate the interpretability methods—SHAP, LIME, and a Hybrid SHAP-LIME ensemble, we designed a systematic experimental framework. Our method- ology centers on consistent modeling, controlled perturbation, and rigorous evaluation metrics across original and noisy datasets.

4.1 Dataset and Preprocessing

We utilize the publicly available LendingClub dataset [10], which contains loan application records from 2007 to 2020. After filtering to include only loans labeled "Fully Paid" or "Charged Off," we created a binary classification target. Preprocessing steps included handling missing values with mean imputation, one-hot encoding for categorical variablesand standardizing continuous variables using z-score normalization. This preprocessing ensures model robustness and fair comparison across explanation techniques.

4.2 Model Training

We trained a Random Forest classifier using an 80-20 train-test split. Hyperparameters were selected through cross- validation. The classifier serves as the consistent black-box model whose predictions are explained using SHAP, LIME, and the Hybrid method.

4.3 Explanation Frameworks

Each explanation method was applied to the same set of test instances under two controlled conditions:

1) **Original Data:** Baseline condition with unaltered test data.
2) **Noisy Data:** Gaussian noise ($\mu = 0$, $\sigma = 0.1$) was added to randomly selected columns in each iteration to emulate real-world perturbations and test explanation robustness.

Each method was executed over 10 independent iterations per test instance to adequately capture variability and facilitate statistical analysis.

4.4 Explanation Methods

- **LIME: [1]** Local surrogate models were fitted using perturbed samples and weighted linear regression. It is model-agnostic and fast but known for unstable explanations.
- **SHAP: [2]** Kernel SHAP was utilized to estimate feature contributions for each instance. Although Kernel SHAP provides theoretical guarantees of consistency and local accuracy, it is computationally intensive, especially in high-dimensional settings or when applied to large datasets. For tree-based models, more efficient alternatives such as Tree SHAP can significantly reduce computational overhead by

leveraging the structure of decision trees. Nonetheless, Kernel SHAP was selected in this study to preserve model-agnostic compatibility across all explanation methods.

- **Hybrid SHAP-LIME:** The Hybrid methodology integrates the complementary advantages of SHAP and LIME through a two-phase process. Initially, SHAP is employed to rank and select the top-J most influential features for a given instance, leveraging its game-theoretic foundation to ensure global attribution consistency. Subsequently, LIME is applied to construct a sparse, locally faithful surrogate model, constrained to this SHAP-informed feature subset. From this local model, a refined selection of the top-K features ($K < J$) are extracted to produce the final explanation. This hierarchical approach enhances interpretability by anchoring local explanations in globally sound feature importance, while also improving explanation stability through reduced perturbation variability.

4.5 *Evaluation Metrics*

We assessed the stability and consistency of the explanations using the following metrics:

- **Spearman Rank Correlation (ρ):** Measures rank-order correlation between feature importances across runs. Higher values indicate greater stability.
- **Standard Deviation of Weights:** Captures the variability in feature attribution weights across iterations. Lower values suggest consistent explanations.
- **Jaccard Stability Index:** Quantifies the intersection- over-union (IoU) of top-K feature sets across runs. This reflects the reproducibility of important features.
- **Kendall Tau (τ):** Evaluates ordinal correlation between ranked lists of features. It complements Spearman by emphasizing order-preserving consistency.

All metrics were averaged over all samples and runs for both original and noisy datasets to ensure statistical robustness. This evaluation framework allows us to distinguish methods based on consistency, resilience to noise, and suitability for sensitive domains like finance.

5. Comparative Evaluation

We report the performance of SHAP, LIME, and Hybrid SHAP-LIME explanations across two settings: on original test data and on data with Gaussian noise added to simulate real- world variability. Each method was run for 10 iterations per example, and the following metrics were averaged across runs: Spearman rank correlation, weight standard deviation, Jaccard stability, and Kendall Tau.

TABLE I PERFORMANCE ON ORIGINAL TEST DATA. HIGHER VALUES OF SPEARMAN, JACCARD, AND KENDALL INDICATE BETTER STABILITY; LOWER STANDARD DEVIATION IS PREFERRED.

Method	Spearman	Weight Std Dev	Jaccard Stability	Kendall Tau
SHAP	**0.902**	**0.001**	**0.681**	**0.855**
LIME	0.673	0.003	0.503	0.600
Hybrid	0.860	**0.001**	0.631	0.793

TABLE II PERFORMANCE ON NOISY DATA (STD = 0.1 NOISE ADDED TO RANDOM FEATURES).

Method	Spearman	Weight Std Dev	Jaccard Stability	Kendall Tau
SHAP	0.813	**0.001**	**0.869**	0.723
LIME	0.673	0.003	0.503	0.600
Hybrid	**0.898**	0.002	0.656	**0.843**

In the clean dataset, SHAP outperforms all methods in cooperative game theory, which ensures additive consistency and robust global-local alignment. SHAP's kernel explainer approximates the contribution of each feature based on marginal expectations, which works especially well when the feature space is not distorted by noise. This leads to high Spearman and Kendall rank correlations, extremely lowvariability in feature attributions, and consistent identification of top features.

However, SHAP's performance deteriorates when Gaussian noise is introduced. The reason lies in its assumption of feature independence and reliance on global sampling distributions. When irrelevant noise enters the data, SHAP can over-attribute importance to perturbed dimensions, reducing fidelity to local model behavior.

LIME, on the other hand, remains relatively unaffected in score by the presence of noise because it always relies on localized sampling and surrogate modeling. across the board. This is largely due to its foundation However, this comes at the cost of high variability and low reproducibility even on clean data. Its inherently stochastic perturbation and sampling make it sensitive to initialization, which translates into low Spearman/Kendall correlations and inconsistent top feature rankings across runs.

The Hybrid SHAP-LIME method shows a compelling trade-off. While it does not surpass SHAP on clean data, it is far more resilient to noise. By leveraging SHAP for top-J feature selection and constraining LIME's perturbation to only relevant dimensions, it limits the influence of irrelevant or noisy features. The addition of Lasso regularization in the local model further enforces sparsity, yielding higher attribution stability and rank agreement. This design enables Hybrid to outperform both SHAP and LIME under noisy conditions while maintaining near-competitive results in cleaner environments.

TABLE III COMPARISON OF SHAP, LIME, AND HYBRID APPROACH

Method	Strengths	Limitations	Best Applied When
LIME	Model-agnostic; intuitive and fast; localized approximation of decision boundary	High variability in explanations; sensitive to perturbations and initialization; unstable across runs	Local feature influence matters; data distribution is stable; explanation speed is crucial
SHAP	Theoretically sound; ensures consistency and additivity; robust on clean data	Computationally intensive; assumes feature independence; susceptible to noise in input space	Interpretability demands are high; reliable attribution needed in low-noise environments
Hybrid	Balances local and global interpretability; improves stability via SHAP-filtered LIME; effective under noisy conditions	Requires careful tuning of feature count and surrogate model; added complexity in design	Data contains noise or feature correlation; trade-off between fidelity and stability is desired

6. Discussion

Performance varied across instances; SHAP performed

especially well in settings with sparse features and low multi- collinearity, where global attribution assumptions held more robustly. Hybrid showed the best consistency when noise or feature interaction complexity increased. Method selection should be context-specific and guided by use-case priorities.

7. Challenges And Future Directions

While local explanation methods such as SHAP, LIME, and hybrid approaches have demonstrated considerable promise, several critical challenges remain. A foremost limitation of this study is that the evaluation was conducted solely on the LendingClub dataset. This restricts the generalizability of the findings to other domains or data distributions. Future work should extend the empirical analysis to a broader range of datasets varying in size, dimensionality, and application domain, to rigorously assess the robustness and adaptability of these explanation techniques.

Moreover, existing evaluation metrics focus primarily on algorithmic consistency and stability, often neglecting the human-centric dimension of interpretability. Incorporating user trust assessments, domain expert validation, and qualitative feedback can offer deeper insights into the real-world usability of explanations. Additionally, advancing explanation frameworks through the integration of causal inference and counterfactual reasoning may improve the actionability and accountability of model insights.

Another key challenge lies in the practical deployment of explanation methods. Computationally intensive techniques like Kernel SHAP may not be suitable for real-time applications or resource-constrained environments. Developing efficient approximations and stream-compatible interpretability tools will be essential for enabling scalable, interpretable AI systems in production-grade settings.

8. Conclusion

We presented a comparative survey of SHAP, LIME, and Hybrid local interpretability methods for credit scoring. Our results show the hybrid approach offers a compelling balance between SHAP's stability and LIME's local modeling. Performance varies by context, and practitioners should align method choice with data properties and regulatory demands.

9. Acknowledgments

The authors would like to express their gratitude for the academic support received during the preparation of this work. No external funding or institutional financial assistance was involved in this study. The authors also declare that there is no conflict of interest related to this article. All data used in this research are fully available within the article, and no additional datasets were generated or sourced externally.

References

1. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
2. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
3. D. Slack, S. Hilgard, E. Jia, S. Singh, and R. Sohoni, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2020, pp. 180–186.
4. D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," arXiv preprint arXiv:1806.08049, 2018.
5. S. Krishna and H. Lakkaraju, "Disagreement among local explanation methods: A comparative study on real-world datasets," in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2022, pp. 678–689.
6. S. Carta et al., "Explainable AI in finance: A survey," arXiv preprint arXiv:2102.01130, 2021.
7. G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," Information Fusion, vol. 76, pp. 89–106, 2021.
8. U. Bhatt et al., "Explainable machine learning in deployment," arXiv preprint arXiv:2011.01962, 2020.
9. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining ma- chine learning classifiers through diverse counterfactual explanations," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT), 2020, pp. 607–617.
10. Lending Club Loan Data, Kaggle Dataset, https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv?resource=download&select=LCDataDictionary.xlsx