Proceedings of the 10th International Conference on Economic Management and Green Development
DOI: 10.54254/2754-1169/2026.35663

Explaining Credit Scoring Models in Digital Lending: A Comparison of SHAP and LIME

Zekun Li

International College Beijing, China Agricultural University, Beijing, China
zekun0905@gmail.com

Abstract. Credit scoring has become more data-driven as online lending platforms collect larger and more varied borrower records. Machine learning methods can model nonlinear patterns that traditional scorecards often miss, but their decisions are harder to explain in a regulated lending environment. This paper discusses how explainable artificial intelligence can be used to make credit scoring models more transparent, with a focus on SHAP and LIME. Using the Lending Club dataset and recent empirical evidence from credit-risk studies, the paper compares the predictive role of ensemble learning models with the interpretive roles of SHAP and LIME. The discussion shows that ensemble methods can provide strong discrimination across public credit datasets, while the usefulness of a model also depends on whether its outputs can be audited and communicated. SHAP is better suited to global feature analysis, model review, and risk-policy design. LIME is more useful when a single loan decision must be explained to staff or customers. Used together, the two methods offer a practical route to balance accuracy, transparency, and compliance in credit scoring.

Keywords: credit scoring, explainable artificial intelligence, SHAP, LIME, credit risk management

1. Introduction

Lending decisions depend heavily on the way borrower risk is measured. In retail credit, the usual goal is to estimate whether a borrower may default by looking at income, debt burden, prior credit use, loan size, and maturity. Older scoring systems usually relied on scorecards or Logistic regression. Their appeal is straightforward: each variable has a visible direction, and loan officers can explain why, for instance, a high debt-to-income ratio or repeated delinquency raises concern.

The data environment has changed with digital lending. Online platforms now process borrower information that is high-dimensional, unevenly distributed, and often nonlinear. A loan amount that appears moderate for one borrower may be risky for another, depending on income, purpose, and credit grade. In the same way, a debt-to-income ratio has to be read alongside past repayment behavior. Machine learning models are useful in this setting because they can capture interactions that linear models simplify. Recent studies show that machine learning can strengthen credit-risk classification [1], and a stacked classifier has reported AUC values of 0.934, 0.944, and 0.870 on the Australian, German, and Taiwan datasets [2].

DOI: 10.54254/2754-1169/2026.35663

Accuracy alone is not enough in this area. Credit scoring affects loan access, pricing, portfolio quality, and the institution's ability to justify its decisions. A model that produces a probability without a usable explanation can create problems in appeal handling, internal validation, and regulatory review. Regulatory guidance also points in this direction: lenders using complex algorithms still need to give consumers specific reasons for adverse credit decisions [3], while complexity, traceability, and limited managerial understanding remain barriers when machine learning is introduced into internal rating models [4].

For this reason, explainable artificial intelligence has become a practical issue rather than only a technical topic. In finance, credit management is one of the common areas where XAI is used, and SHAP and LIME are among the most widely adopted post-hoc explanation tools [5]. Interpretability in credit scoring should include transparency and auditability, not only a list of important variables [6]. This paper asks three questions: how machine learning improves default identification, what SHAP and LIME are best able to explain, and how their outputs can support usable risk-management decisions.

2. Research design

2.1. Data source and research object

The empirical context of the paper is the public Lending Club loan data, a commonly used source in credit-scoring research. The data contain borrower attributes, loan information, credit history, loan purpose, and the final repayment outcome. The credit-granting dataset prepared by Ariza-Garzón et al. covers loans issued from 2007 to 2018 [7]. After loans with unresolved transitional statuses were removed, the dataset contained 1,347,681 observations. In that dataset, 'Charged Off' is treated as default and 'Fully Paid' as normal repayment.

The feature set should reflect what a lender would know before approval. Variables created after the loan has been issued, such as final payment date, collection status, or actual recovery amount, should be excluded. If such variables enter the model, the reported performance may look strong, but the result would rely on data leakage and would not represent a real pre-loan decision process.

2.2. Variables and preprocessing

The candidate variables can be grouped into repayment capacity, credit history, loan characteristics, and borrower stability. Annual income, debt-to-income ratio, and loan amount describe repayment capacity. FICO score, past delinquency, and recent inquiries reflect credit history. Loan term, purpose, and interest rate describe the credit product. Employment length and home ownership provide information on borrower stability. Together, these variables describe ability to repay, willingness to repay, and short-term pressure on liquidity.

Preprocessing is needed before modelling. Missing numerical values may be replaced with medians, while missing categories can be coded as unknown. Nominal variables such as loan purpose and home ownership can be converted through one-hot encoding, whereas ordered credit grades can be label-coded. Because default cases are usually fewer than fully repaid loans, the training process may use class weights, threshold adjustment, or SMOTE to reduce majority-class bias. Class imbalance is an important issue in credit-risk classification [2].

DOI: 10.54254/2754-1169/2026.35663

2.3. Models and evaluation metrics

This paper treats Logistic regression as the traditional baseline and discusses random forest, XGBoost, and LightGBM as representative machine learning models. Random forest reduces variance by aggregating many decision trees. XGBoost and LightGBM are gradient-boosting methods that are widely used for structured tabular data. In a bank consumer-loan setting, LightGBM combined with SHAP can offer both predictive performance and interpretable variable contributions [8].

Model evaluation should reflect the risk objective. Accuracy may be misleading when defaults form only a small share of the sample. The paper therefore considers AUC, F1-score, recall, and KS. AUC measures how well a model separates good and bad borrowers; F1-score balances precision and recall; recall focuses on the ability to capture default cases; and KS measures the largest distance between the cumulative distributions of good and bad customers.

2.4. SHAP and LIME

SHAP explains a prediction by assigning each feature a contribution value. A positive value pushes the prediction toward higher risk, while a negative value lowers it. The same framework can be aggregated to show the variables that matter most across the portfolio, or used locally to explain one borrower. Prior work has shown that SHAP can explain both local and overall outputs of deep learning credit-scoring models [9].

LIME takes a different route. It creates small perturbations around the borrower being explained, observes how the black-box model reacts, and fits a simple local surrogate model to that small neighborhood. The output is easy to read and is often useful when an individual decision has to be discussed with staff or a customer. LIME has been applied to credit-scoring models, with reported accuracy values of 88.84%, 78.30%, and 77.80% on the Australian, German, and South German datasets [10].

3. Results and analysis

3.1. Prediction performance

Recent evidence suggests that ensemble methods often perform better than single classifiers on public credit datasets. A credit-risk prediction engine using information gain for feature selection and a stacked classifier reports AUC values of 0.934 for the Australian dataset, 0.944 for the German dataset, and 0.870 for the Taiwan dataset [2].

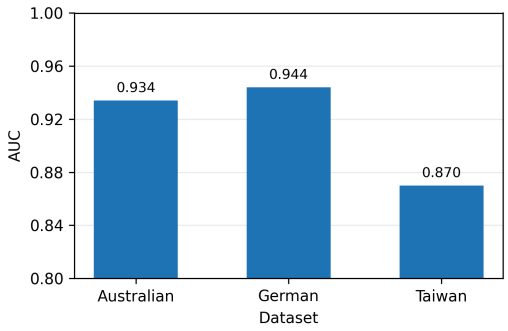

Figure 1. AUC performance of the stacked classifier on three credit datasets

DOI: 10.54254/2754-1169/2026.35663

The German dataset has the highest AUC, followed by the Australian and Taiwan datasets. All three values exceed 0.85, which suggests that the stacked classifier has meaningful discrimination ability across different credit data settings. At the same time, the difference between 0.944 and 0.870 is not trivial. It shows that sample composition, variable quality, and class imbalance still affect model behavior even when the same modelling strategy is used.

The same study also reports Accuracy and F1-score [2]. The Australian dataset records 86.23% Accuracy and an F1-score of 84.58%. The German dataset records 82.80% and 86.35%. The Taiwan dataset records 85.80% Accuracy but only 51.35% F1-score. This contrast is important for credit scoring because a model may appear accurate while still missing many default cases.

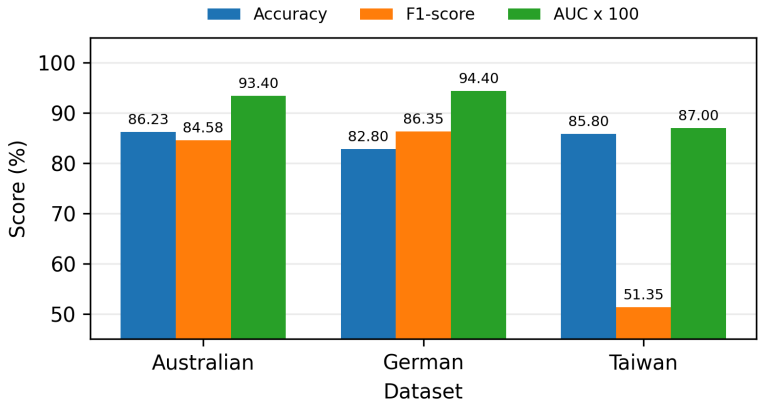

Figure 2. Accuracy, F1-score, and AUC of the stacked classifier

Figure 2 makes this point more visible. The Australian and German datasets show a more balanced pattern across Accuracy, F1-score, and AUC. By contrast, the Taiwan dataset shows a large gap between Accuracy and F1-score. In practical lending, such a result would require further threshold review or imbalance treatment before the model could be used for approval decisions.

3.2. Global explanation value of SHAP

In credit scoring, SHAP is most useful when a lender needs to know which factors drive model behavior at the portfolio level. A complex model such as XGBoost or LightGBM may contain many trees and splits, so reading the model structure directly is not practical. SHAP translates the model output into feature contributions that can be ranked and inspected. Studies use SHAP to explain LightGBM in credit assessment [8] and describe SHAP as a tool that can improve transparency in automated credit decisions [11].

The important variables are also familiar to credit practitioners. Debt-to-income ratio, FICO score, loan amount, loan term, past delinquency, revolving utilization, and annual income often appear as major risk signals. A high debt-to-income ratio points to repayment pressure, a low FICO score signals weaker historical credit performance, and high revolving utilization may suggest short-term liquidity strain. If these variables appear near the top of a SHAP ranking, the model is at least learning a pattern that aligns with lending practice.

SHAP is also useful because it can show nonlinear effects. An increase in credit score may reduce risk sharply at the lower end of the distribution, while adding fewer benefits once a borrower is already in a high-score range. Debt-to-income ratio may behave in a similar way: below a certain range it may add limited risk, but above a threshold it can raise default probability quickly. The

DOI: 10.54254/2754-1169/2026.35663

reliability of the explanation attached to machine learning credit decisions should therefore be considered alongside predictive quality [12].

3.3. Local explanation value of LIME

LIME is better suited to a narrower question: why was this particular borrower treated as risky or safe? It builds an explanation around the case being reviewed. By fitting a simple model around nearby perturbed examples, it provides a local account of the decision. This makes the output closer to the language used in loan review.

Suppose a borrower is classified as high risk. A LIME explanation may show that the decision is mainly pushed by a high debt-to-income ratio, several recent inquiries, earlier delinquency, and a longer loan term. It may also identify factors that partly reduce risk, such as stable employment or higher income. This type of explanation is more useful than simply saying that the system score was too low, because it gives staff a defensible way to discuss the case.

The weakness of LIME is that the explanation depends on the local sampling design. Changes in perturbation settings, neighborhood size, or feature scaling may change the weights. Both SHAP and LIME can be affected by model type and feature correlation [13]. For credit data, where many variables are related to one another, explanation stability should therefore be checked rather than assumed.

3.4. Comparison between SHAP and LIME

SHAP and LIME should not be viewed as competing tools. They answer different operational questions. SHAP is better for asking how the model works overall, which variables dominate, and whether the risk logic is plausible. LIME is better for explaining a specific approval or rejection decision. The first is more valuable to model validation and risk teams; the second is more useful to loan officers and customer-facing processes.

Table 1. Operational differences between SHAP and LIME in credit scoring

Dimension	SHAP	LIME	Financial use
Scope	Portfolio-level and case-level explanations.	Mainly case-level explanations.	SHAP supports validation; LIME supports case review.
Output	Feature contribution values and importance ranking.	Weights from a local surrogate model.	SHAP helps policy design; LIME helps communication.
Stability	Generally more stable, but still affected by feature correlation.	Sensitive to sampling, neighborhood size, and scaling.	Both require stability checks before use.
Use case	Model development, monitoring, and audit.	Approval review and adverse-decision explanation.	Using both methods is more practical than using either alone.

Note. The comparison is organized according to the mechanisms of SHAP and LIME and their use in credit-scoring workflows.

Table 1 summarizes the difference. SHAP has stronger value in global feature ranking and model review, although it may still be influenced by correlated variables. LIME is easier to understand in a single case, but its local explanation can move with the perturbation design. Explainable AI in financial decision support should be assessed in relation to business usability, not only technical interpretability [14].

DOI: 10.54254/2754-1169/2026.35663

A practical credit-scoring workflow can therefore use both. SHAP can be applied during model development and periodic validation to check whether the model depends on reasonable signals. LIME can be used later when a case requires a brief explanation for review or customer communication. This division avoids relying on one method for every task.

4. Financial applications and discussion

In pre-loan approval, explanations help a lender move beyond a pass-or-fail score. A high-risk decision may be driven by short-term repayment pressure, weak credit history, or a loan amount that does not match the borrower's income. These sources of risk imply different actions. The lender may reduce the credit limit, shorten the maturity, tighten approval standards, or adjust pricing. Without explanation, these choices are harder to justify.

In post-loan management, borrower risk changes after the loan is issued. Income, debt, repayment behavior, and credit use may all shift over time. If a risk score rises and SHAP shows that the change is linked to higher revolving utilization and more recent inquiries, the signal may indicate liquidity pressure. If the increase is driven by repeated delinquency, the issue is more directly connected with repayment behavior.

Robustness checks are also needed. A model should be tested outside the original training sample and across customer groups or time periods. Cross-model consistency is another useful check. If XGBoost, LightGBM, and random forest all rank debt-to-income ratio, FICO score, and delinquency history as important, those variables are more credible as risk factors. Top-ranked SHAP features are usually more stable than variables with medium importance [15].

Fairness remains a separate concern. Even when sensitive variables are not used directly, proxies such as location, income structure, or job stability can create uneven effects. Transparency, fairness, and predictive performance can be considered together in automated credit decisions [11]. For that reason, lenders should keep records of data sources, compare complex models with a baseline, and document SHAP and LIME explanations when decisions are reviewed.

5. Conclusion

This paper examines the use of explainable artificial intelligence in credit scoring. Drawing on public Lending Club data and recent credit-risk studies, it argues that ensemble models such as random forest, XGBoost, LightGBM, and stacked classifiers can handle nonlinear relationships more effectively than traditional linear models. Yet the value of a credit-scoring model depends not only on AUC, Accuracy, or F1-score. Because lending decisions affect customers and institutions directly, the model must also be understandable and reviewable.

SHAP and LIME contribute to that goal in different ways. SHAP is better suited to global model explanation, feature ranking, and model governance. LIME is more useful for a local explanation of a single borrower case. Used together, they provide a workable framework in which SHAP supports validation and policy design, while LIME supports individual case review and communication.

The paper has limitations. The data are drawn mainly from the U.S. online lending market, so the findings may not transfer fully to domestic consumer finance settings. The comparison also focuses on SHAP and LIME, while counterfactual and causal explanation methods are not examined in detail. Future work could use desensitized domestic financial data and compare a wider set of explanation methods, especially where customer rights and fairness need to be evaluated quantitatively.

DOI: 10.54254/2754-1169/2026.35663

References

[1] Suhadolnik, N., Ueyama, J., & da Silva, S. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management, 16*(12), 496. https://doi.org/10.3390/jrfm16120496

[2] Ileberi, E., Sun, Y., & Wang, Z. (2024). A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data, 11*, 23. https://doi.org/10.1186/s40537-024-00882-0

[3] Consumer Financial Protection Bureau. (2023). Circular 2023-03: Adverse action notification requirements and the proper use of the CFPB's sample forms provided in Regulation B. https://files.consumerfinance.gov/f/documents/cfpb_adverse_action_notice_circular_2023-09.pdf

[4] European Banking Authority. (2023). Follow-up report on machine learning for IRB models. https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf

[5] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review, 57*, 216. https://doi.org/10.1007/s10462-024-10854-8

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

[7] Ariza-Garzón, M. J., Sanz-Guerrero, M., Arroyo Gallardo, J., & Lending Club. (2024). Lending Club loan dataset for granting models (Version 0.1) [Data set]. Universidad Complutense de Madrid. https://doi.org/10.5281/zenodo.11295916

[8] de Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for credit assessment in banks. *Journal of Risk and Financial Management, 15*(12), 556. https://doi.org/10.3390/jrfm15120556

[9] Hjelkrem, L. O., & de Lange, P. E. (2023). Explaining deep learning models for credit scoring with SHAP: A case study using open banking data. *Journal of Risk and Financial Management, 16*(4), 221. https://doi.org/10.3390/jrfm16040221

[10] Aljadani, A., Alharthi, B., Farsi, M. A., Balaha, H. M., Badawy, M., & Elhosseini, M. A. (2023). Mathematical modeling and analysis of credit scoring using the LIME explainer: A comprehensive approach. *Mathematics, 11*(19), 4055. https://doi.org/10.3390/math11194055

[11] Nwafor, C. N., Nwafor, O., & Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach. *Scientific Reports, 14*, 25174. https://doi.org/10.1038/s41598-024-75026-8

[12] Alonso-Robisco, A., & Carbó, J. M. (2025). Should we trust the credit decisions provided by machine learning models? *Computational Economics, 66*(5), 4245–4274. https://doi.org/10.1007/s10614-025-10855-x

[13] Salih, A. M., Raisi-Estabragh, Z., Boscolo Galazzo, I., Radeva, P., Petersen, S. E., Menegaz, G., & Lekadir, K. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems, 7*, 2400304. https://doi.org/10.1002/aisy.202400304

[14] Nallakaruppan, M. K., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V. P., & Hameed, I. A. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence. *Risks, 12*(10), 164. https://doi.org/10.3390/risks12100164

[15] Lin, L., & Wang, Y. (2025). SHAP stability in credit risk management: A case study in credit card default model. *Risks, 13*(12), 238. https://doi.org/10.3390/risks13120238