International Journal of Computer Techniques–IJCT Volume 13 Issue 3, May-June - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

RESEARCH ARTICLE OPEN ACCESS

Explainable Credit Risk Assessment: Comparing Logistic Regression, XGBoost, and LightGBM with SHAP and LIME Analysis

Durrusadaf Imanova*, Hakan Kutucu**

* (Azerbaijan State University of Economics, Azerbaijan

Email: imanovadurrusdf@gmail.com)

** (Department of Software Engineering, Karabuk University, Türkiye)

----------**********----------

Abstract:

This study investigates whether gradient boosting models (XGBoost and LightGBM), combined with SHAP and LIME explainability, offer meaningful advantages over Logistic Regression in credit scoring. Using the German Credit dataset (n = 1,000), three models were compared through stratified five-fold cross-validation, held-out test evaluation, and pairwise McNemar's tests. None of the three models achieved a statistically significant improvement over any other (all $p > 0.05$). Logistic Regression achieved the highest recall on the test set (0.800), followed by XGBoost (0.717) and LightGBM (0.567). SHAP analysis revealed consistent nonlinear risk patterns across both ensemble models, including duration threshold effects and compensatory attribute dynamics inaccessible through linear coefficients. A comparative analysis of SHAP and LIME showed broad agreement on top predictors but meaningful differences in stability and explanation structure. The findings suggest that on small, structured datasets, Logistic Regression remains the most practical production choice, while gradient boosting with SHAP serves a complementary analytical role for credit policy design.

Keywords — **Explainable AI, Credit Risk, Logistic Regression, XGBoost, LightGBM, SHAP, LIME**

----------**********----------

I. INTRODUCTION

Credit risk assessment is a fundamental component of banking operations. Financial institutions must continuously evaluate borrower repayment likelihood, as inaccurate credit decisions result in loan defaults, capital inefficiency, and reputational damage. Logistic Regression has dominated credit scoring for decades due to its statistical transparency and regulatory compatibility.

However, modern lending datasets are increasingly complex. Machine learning algorithms, particularly XGBoost [2] and LightGBM [8], have achieved strong predictive performance on structured tabular data but are considerably more difficult to interpret. Explainable AI (XAI) attempts to bridge this gap through post-hoc explanation methods such as SHAP [10] and LIME [12].

While numerous studies have compared machine learning with traditional scoring methods, most focus on demonstrating predictive superiority. Fewer studies examine whether that superiority persists on small datasets, compare multiple XAI methods head-to-head, or assess whether

explainability tools carry independent analytical value when performance gains are limited.

This paper compares Logistic Regression, XGBoost, and LightGBM within a unified explainable credit scoring framework using both SHAP and LIME. The study addresses four research questions: (1) Do ensemble methods outperform Logistic Regression on a small structured credit dataset? (2) Do XGBoost and LightGBM learn similar risk structures? (3) Do SHAP and LIME provide consistent explanations? (4) Which modeling strategy is most practical under transparency requirements?

II. LITERATURE REVIEW

Credit scoring has relied on statistical classification methods since foundational surveys by Hand and Henley [7] and Thomas [13] established Logistic Regression as the industry standard. Large-scale benchmarking by Baesens et al. [1] found performance differences between models were often small, while Lessmann et al. [9] concluded that ensemble methods can achieve significant improvements on larger datasets but that the gap narrows on smaller ones. Crook et al. [3] documented growing adoption of nonparametric methods alongside practical barriers in regulated environments.

Among ensemble methods, XGBoost [2] uses regularised tree boosting, while LightGBM [8] employs leaf-wise growth and histogram-based splitting for faster training. Both have become standard benchmarks in credit scoring research.

SHAP [10], grounded in cooperative game theory, provides theoretically consistent feature attributions. LIME [12] generates local surrogate models by perturbing inputs around each prediction. De Lange et al. [4] applied SHAP to credit scoring on Norwegian bank data; Hadji Misheva et al. [6] argued SHAP enhances regulatory compliance. However, direct SHAP-LIME comparisons in credit contexts remain rare. The present study addresses this gap by comparing three models with both explanation methods.

III. DATA AND METHODOLOGY

A. Dataset and Preprocessing

The German Credit dataset [5] contains 1,000 loan applications with 20 borrower attributes and a binary target: good credit (n = 700) or bad credit (n = 300). Thirteen categorical features were one-hot encoded (drop_first=True), expanding to 48 columns. An 80/20 stratified split produced training (n = 800) and test (n = 200) subsets. Standardization was applied only to Logistic Regression inputs.

B. Model Development

Logistic Regression was trained with L2 regularization (C = 1.0) and class_weight = 'balanced'. The model converged in 26 iterations.

XGBoost was configured with scale_pos_weight = 2.33. Grid search over 324 combinations (max_depth {3,4,5,6}, learning_rate {0.01,0.05,0.1}, n_estimators {100,150,200}, subsample {0.7,0.8,0.9}) with stratified 5-fold CV optimizing ROC-AUC yielded: max_depth = 3, learning_rate = 0.05, n_estimators = 200, subsample = 0.8 (CV AUC = 0.792).

LightGBM was configured with scale_pos_weight = 2.33 and is_unbalance = False. Grid search over 108 combinations (num_leaves {20,31,50}, learning_rate {0.01,0.05,0.1}, n_estimators {100,150,200}, min_child_samples {10,20,30}) yielded: num_leaves = 20, learning_rate = 0.05, n_estimators = 150, min_child_samples = 20 (CV AUC = 0.786). The small optimal num_leaves confirms limited capacity to exploit complex structures at this sample size.

C. Evaluation and Explainability

Performance was assessed via stratified 5-fold CV, held-out test evaluation, and pairwise McNemar's tests across all three model pairs. Recall and ROC-AUC were given primary interpretive weight due to the cost asymmetry between false negatives and false positives in lending.

For explainability, Logistic Regression coefficients were interpreted directly. For both ensemble models, SHAP TreeExplainer produced global importance rankings, summary plots, dependence plots, and individual waterfall

explanations. LIME (LimeTabularExplainer) was applied to the same representative predictions to enable direct SHAP-LIME comparison on consistency, stability, and explanation structure.

IV. EMPIRICAL RESULTS

A. Classification Performance

Table I presents cross-validation results. All three models achieve comparable performance, with differences within one standard deviation. LightGBM achieves the highest accuracy and precision but the lowest recall (0.607), reflecting a tendency to favour specificity over sensitivity under balanced weighting.

TABLE I. STRATIFIED 5-FOLD CROSS-VALIDATION RESULTS

Metric	Logistic Regression	XGBoost	LightGBM
Accuracy	0.724 ± 0.023	0.729 ± 0.017	0.734 ± 0.034
Precision	0.530 ± 0.029	0.538 ± 0.022	0.554 ± 0.052
Recall	0.723 ± 0.034	0.697 ± 0.025	0.607 ± 0.043
F1-score	0.611 ± 0.029	0.607 ± 0.020	0.579 ± 0.046
ROC-AUC	0.781 ± 0.021	0.787 ± 0.025	0.770 ± 0.031

Table II reports test set results. Logistic Regression leads on all five metrics, with a substantial recall advantage (0.800 vs. 0.717 and 0.567). Moderate precision (~0.53–0.56) reflects the balanced class weighting strategy prioritising default detection. LightGBM achieves the lowest recall (0.567), missing 26 of 60 defaulters, making it the least suitable for default-detection objectives.

TABLE II. TEST SET RESULTS (POSITIVE CLASS: BAD CREDIT)

Metric	Logistic Regression	XGBoost	LightGBM
Accuracy	0.750	0.725	0.720
Precision	0.558	0.531	0.531
Recall	0.800	0.717	0.567
F1-score	0.658	0.610	0.548
ROC-AUC	0.803	0.795	0.788

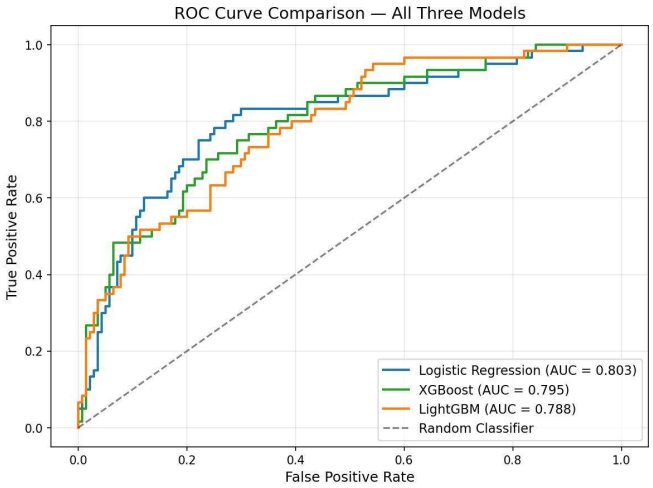

Fig. 1. ROC Curve Comparison — All Three Models

The confusion matrices (Fig. 2) reveal that all three models correctly classify 102–110 of 140 good credit cases. The critical difference lies in the bad credit class: Logistic Regression identifies 48 of 60 defaulters, XGBoost 43, and LightGBM only 34.

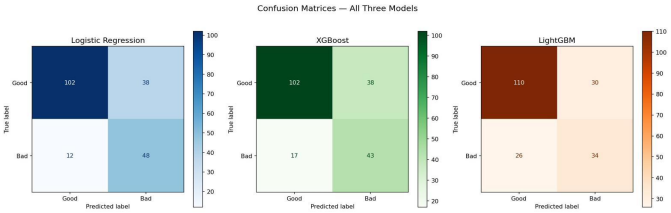

Fig. 2. Confusion Matrices — Logistic Regression, XGBoost, and LightGBM

B. Statistical Significance

McNemar's tests across all three pairs confirm no statistically significant differences: LR vs. XGBoost ($\chi^2 = 0.552$, $p = 0.458$), LR vs. LightGBM ($\chi^2 = 0.595$, $p = 0.440$), and XGBoost vs. LightGBM ($\chi^2 = 0.000$, $p = 1.000$). No model is statistically superior on this dataset.

C. SHAP Explainability Analysis

Global feature importance. SHAP analysis of both ensemble models reveals consistent importance rankings. Checking account A14 dominates in both XGBoost (mean |SHAP| = 0.726) and LightGBM (1.03), followed by duration and credit amount (Fig. 3). This cross-model consistency strengthens confidence in the identified risk drivers.

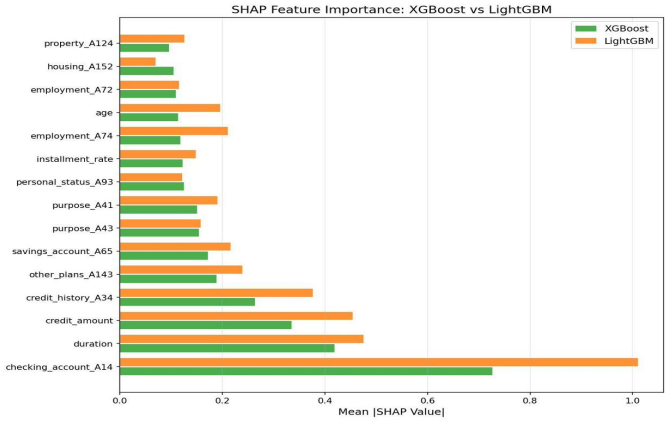

Fig. 3. SHAP Feature Importance — XGBoost vs. LightGBM

Nonlinear patterns. SHAP dependence plots (Fig. 4) reveal that credit duration shows gradual risk increases up to ~30 months, then accelerates steeply. Credit amount displays similar threshold behaviour. These dynamics are structurally inaccessible under Logistic Regression's linear framework.

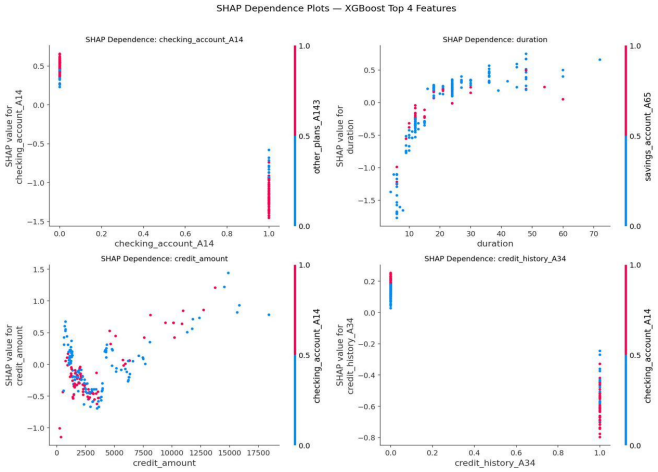

Fig. 4. SHAP Dependence Plots — Top 4 Features (XGBoost)

Individual explanations. For a bad credit case (P(default) = 0.841, Fig. 5), the prediction was driven by absence of a checking account (+0.43), no other plans (+0.39), and 24-month duration (+0.30). For a good credit case (P(default) = 0.388), savings account A64 contributed −1.12, single-handedly offsetting all risk factors — a compensatory dynamic constrained under Logistic Regression's additive structure.

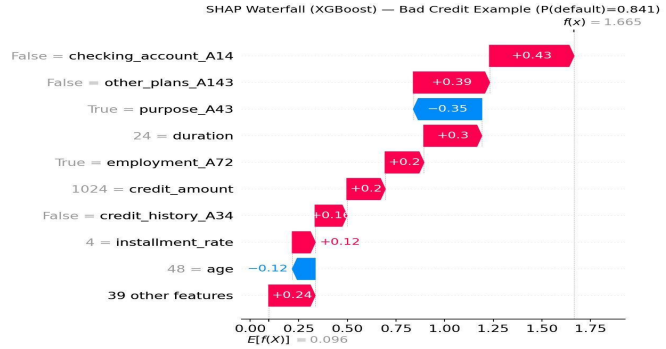

Fig. 5. SHAP Waterfall — Bad Credit Example (P(default) = 0.841)

D. SHAP vs. LIME Comparison

LIME was applied to the same representative predictions used in SHAP analysis. Fig. 6 presents a side-by-side comparison for the bad credit case on XGBoost. Both methods identify checking_account_A14 as a top predictor, but overall only 2 of 5 top features overlap. SHAP assigns the highest contribution to checking_account_A14 and other_plans_A143, while LIME emphasises purpose_A41 and savings account categories.

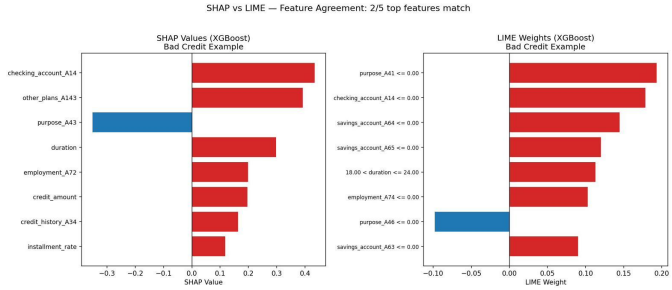

Fig. 6. SHAP vs. LIME Feature Attribution — Bad Credit Example (XGBoost)

A stability analysis running LIME 10 times on the same instance (Fig. 7) shows moderate variability in feature weights, with coefficients of variation ranging from 5–15% for top features. SHAP, being deterministic for tree models via TreeExplainer, produces identical results across runs. This stability advantage makes SHAP more suitable for regulatory audit documentation, while LIME's interpretable linear format may be preferred for customer-facing explanations where approximate consistency is acceptable.

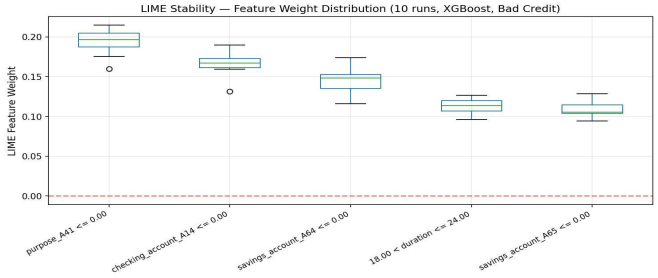

Fig. 7. LIME Stability — Feature Weight Distribution Over 10 Runs (XGBoost, Bad Credit)

V. DISCUSSION

The results challenge the assumption that ensemble methods consistently outperform traditional approaches. None of the three models achieved a statistically significant advantage (all $p > 0.05$). This aligns with Lessmann et al. [9] regarding the limited advantage of complex models on small datasets. Both XGBoost and LightGBM favoured conservative hyperparameters (shallow depth, low learning rate), reflecting limited capacity to exploit nonlinear structures with only 800 training samples.

LightGBM's lower recall (0.567 vs. 0.800 for LR) is particularly noteworthy. Its leaf-wise growth strategy, while efficient on large datasets, did not translate to superior performance here. For lending applications where default detection is the priority, this recall deficit makes LightGBM the least suitable option despite competitive ROC-AUC.

Despite comparable aggregate performance, SHAP analysis reveals structural insights inaccessible through linear coefficients. Both ensemble models consistently identify the same top risk drivers (checking account, duration, credit amount), strengthening confidence in these findings. The duration threshold effect and compensatory savings dynamics inform credit policy design independently of classification accuracy.

The SHAP-LIME comparison reveals a practical trade-off: SHAP provides deterministic, theoretically grounded attributions suitable for audit and compliance; LIME produces intuitive local linear approximations but with measurable instability across runs. For regulatory environments requiring reproducible documentation, SHAP is the stronger choice. For real-time customer-facing explanations, LIME's simpler output format may be acceptable despite its variability.

VI. LIMITATIONS

The German Credit dataset (1,000 observations, 1990s) limits ensemble methods' capacity and contemporary relevance. Findings are based on a single benchmark. Macroeconomic indicators were excluded. Fairness evaluation across demographic groups was not conducted. Class imbalance was addressed solely through class weighting; alternatives such as SMOTE were not explored. LIME stability was assessed on a single instance; broader stability analysis across multiple predictions would strengthen the comparison.

VII. CONCLUSION

This study compared Logistic Regression, XGBoost, and LightGBM within an explainable credit scoring framework using both SHAP and LIME. No model achieved a statistically significant predictive advantage. Logistic Regression achieved the highest recall (0.800) and remains the most practical production choice for small structured portfolios due to its native interpretability and regulatory compatibility.

SHAP analysis retains analytical value independently of model superiority, revealing consistent nonlinear risk patterns across both ensemble models. The SHAP-LIME comparison demonstrates that explanation methods offer complementary rather than redundant perspectives, with SHAP better suited for audit documentation and LIME for interpretable customer communication. Future research should extend this comparison to larger real-world datasets, incorporate macroeconomic variables, evaluate fairness, and examine interpretable ensemble architectures such as Explainable Boosting Machines.

REFERENCES

[1] 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.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD, 2016, pp. 785–794.

[3] J. N. Crook, D. B. Edelman, and L. C. Thomas, "Recent developments in consumer credit risk assessment," Eur. J. Oper. Res., vol. 183, no. 3, pp. 1447–1465, 2007.

[4] P. E. de Lange, B. Melsom, C. B. Vennerød, and S. Westgaard, "Explainable AI for credit assessment in banks," J. Risk Financial Manag., vol. 15, no. 12, p. 556, 2022.

[5] D. Dua and C. Graff, "UCI Machine Learning Repository," Univ. California, Irvine, 2019.

[6] B. Hadji Misheva, J. Osterrieder, A. Hirsa, O. Kulkarni, and S. F. Lin, "Explainable AI in credit risk management," arXiv:2103.00949, 2021.

[7] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring," J. Roy. Stat. Soc. A, vol. 160, no. 3, pp. 523–541, 1997.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 3146–3154.

[9] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update," Eur. J. Oper. Res., vol. 247, no. 1, pp. 124–136, 2015.

[10] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 4765–4774.

[11] C. Molnar, Interpretable Machine Learning, 2nd ed. christophm.github.io/interpretable-ml-book, 2022.

[12] 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.

[13] L. C. Thomas, "A survey of credit and behavioural scoring," Int. J. Forecasting, vol. 16, no. 2, pp. 149–172, 2000.