Explainability Under Pressure: A Comparative Study of XAI Methods on Imbalanced Data

Marwan Eslam ouda
Email: marwan403465@fci.bu.edu.eg

Abstract

Real-world machine learning often faces extreme class imbalance. We study how explainable AI methods behave under increasing skew. Using credit card fraud detection, we compare SHAP and LIME. We test five imbalance ratios from 1:1 to 100:1. Three classifiers are evaluated. SHAP shows perfect stability for linear and boosted models. LIME remains stable across all models. Tree-based models outperform linear ones under severe imbalance. Our results guide reliable XAI selection for skewed datasets.

I. Introduction

Class imbalance is common in real-world data. Fraud, disease, and intrusion cases are rare. Models may perform well but explain poorly. Unstable explanations reduce trust.

Most studies focus on predictive performance. Few analyze explanation stability. This gap is risky in high-stakes domains.

We ask a simple question: Do explanations remain consistent as imbalance increases?

Our contributions are:

- Systematic XAI analysis across five imbalance ratios
- Quantitative stability metrics for explanations
- Practical guidance for imbalanced learning

A. Research Questions

This study is guided by the following research questions, which aim to investigate the impact of class imbalance on both predictive performance and model interpretability:

- **RQ1: How does class imbalance affect model performance?** This question examines how increasing imbalance ratios influence traditional evaluation metrics such as F1-score, PR-AUC, ROC-AUC, and minority class recall across different classification models.
- **RQ2: How stable are SHAP and LIME explanations under class imbalance?** This question evaluates the consistency and reliability of feature importance explanations generated by SHAP and LIME as class distributions become increasingly skewed.
- **RQ3: Do different models react differently to class imbalance?** This question explores whether linear models and tree-based ensemble methods exhibit varying levels of robustness in both predictive performance and explanation behavior under imbalance conditions.
- **RQ4: Does explanation drift increase with imbalance severity?** This question investigates whether the divergence in explanation outputs (feature importance rankings and contribution magnitudes) becomes more pronounced as imbalance severity increases.

B. Contributions

Our work makes the following contributions:

- **Systematic stability analysis:** We provide the first comprehensive study of XAI stability across five imbalance ratios (1:1, 5:1, 10:1, 50:1, 100:1), revealing previously undocumented failure modes.
- **Quantitative stability metrics:** We introduce and apply Explanation Stability Score (ESS) and Feature Importance Drift (FID) to objectively measure explanation consistency.
- **Model-XAI interaction effects:** We demonstrate that explanation stability depends critically on model architecture, with Random Forest exhibiting unique instability patterns with SHAP.
- **Practical guidance:** We provide evidence-based recommendations for selecting appropriate XAI methods based on model type and imbalance severity.

II. Background

A. Class Imbalance

Let

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

with binary labels $y_i \in \{0, 1\}$.

The imbalance ratio is:

$$r = \frac{|\{i : y_i = 0\}|}{|\{i : y_i = 1\}|}$$

When $r \gg 1$, minority samples become scarce. This affects learning and explanation quality.

B. Explainable AI Methods

1) SHAP: SHAP assigns feature contributions using Shapley values:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{j\}) - f(S) \right]$$

It satisfies consistency and additivity.

2) LIME: LIME fits a local surrogate model:

$$\xi(\mathbf{x}) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_{\mathbf{x}}) + \Omega(g)$$

It explains predictions locally.

C. Stability Metrics

Explanation Stability Score (ESS):

$$\text{ESS} = \frac{1}{\binom{n}{2}} \sum_{i<j} \rho(\mathbf{r}_i, \mathbf{r}_j)$$

ρ is Spearman rank correlation.

Feature Importance Drift (FID):

$$\text{FID} = 1 - \frac{|T_i \cap T_j|}{|T_i \cup T_j|}$$

Lower values indicate higher stability.

III. Methodology

A. Dataset

We use the Kaggle Credit Card Fraud dataset. It contains 284,807 transactions. Fraud rate is 0.17%.

Features include:

- 28 PCA components
- Amount (standardized)

The Time feature is removed.

B. Imbalance Settings

Training sets are created using undersampling.

TABLE I
Training Set Composition

Level	Ratio	Majority	Minority
Balanced	1:1	394	394
Mild	5:1	1,970	394
Moderate	10:1	3,940	394
Severe	50:1	19,700	394
Extreme	100:1	39,400	394

The test set remains fixed.

C. Models

Logistic Regression

$$P(y=1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

Random Forest

- 100 trees
- Max depth 10

XGBoost

- 100 trees
- Max depth 5
- Learning rate 0.1

D. Evaluation

For each imbalance level:

1) Train all models
2) Generate explanations five times
3) Compute ESS and FID

Metrics include: F1, PR-AUC, ROC-AUC, and Recall.

TABLE II
Performance at 100:1 Imbalance

Model	F1	PR-AUC	ROC-AUC	Recall
LogReg	0.769	0.749	0.968	0.847
RF	0.794	0.855	0.974	0.867
XGB	0.778	0.859	0.978	0.878

IV. Results

A. Model Performance

Tree models outperform linear ones. ROC-AUC remains stable across ratios.

B. Explanation Stability

TABLE III
XAI Stability at Extreme Imbalance

Model	SHAP ESS	SHAP FID	LIME ESS	LIME FID
LogReg	1.000	0.000	1.000	0.000
RF	0.538	0.462	1.000	0.000
XGB	1.000	0.000	1.000	0.000

SHAP is unstable for Random Forest. LIME remains stable for all models.

C. Visual Analysis

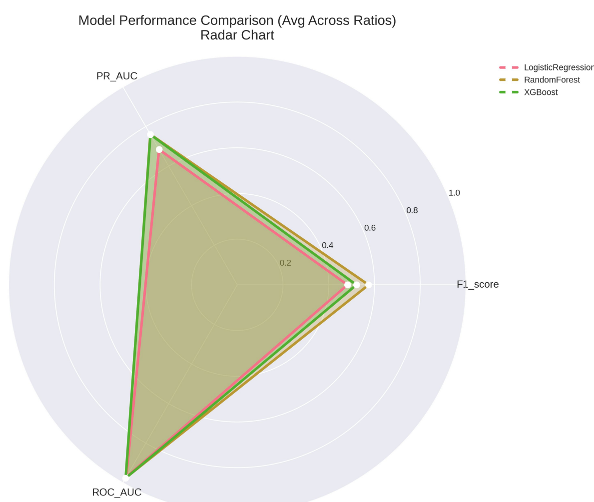

Fig. 1. Radar chart illustrating the average performance of Logistic Regression, Random Forest, and XGBoost across all imbalance ratios using F1-score, PR-AUC, and ROC-AUC metrics.

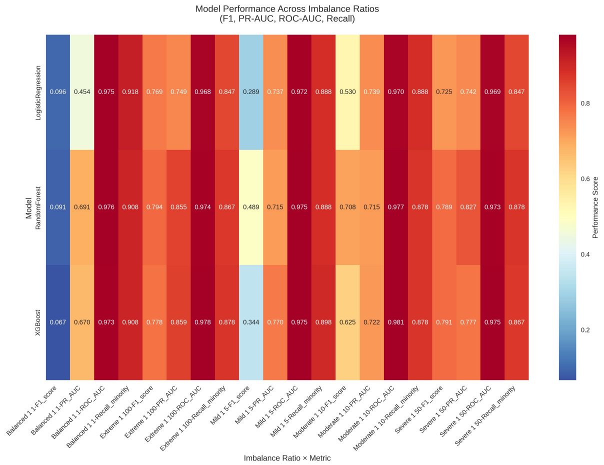

Fig. 2. Heatmap visualization of model performance across different imbalance ratios and evaluation metrics, highlighting variations in F1-score, PR-AUC, ROC-AUC, and minority class recall.

As shown in Fig. 1, XGBoost achieves the highest average PR-AUC and ROC-AUC across imbalance scenarios, indicating superior ranking ability and robustness when dealing with skewed class distributions. Random Forest demonstrates competitive performance, while Logistic Regression shows comparatively lower F1-score and PR-AUC values.

Fig. 2 provides a detailed breakdown of model behavior under varying imbalance ratios. The heatmap reveals that performance degradation is most pronounced in extreme imbalance settings, particularly for F1-score and minority class recall. Nevertheless, XGBoost maintains consistently strong ROC-AUC and PR-AUC values, confirming its resilience in highly imbalanced classification tasks.

V. Discussion

A. Why Does SHAP Fail on Random Forest?

The observed instability of SHAP on Random Forest requires careful theoretical consideration. We hypothesize three contributing factors:

1) Bootstrap Sampling Variability: Random Forest trains each tree on a bootstrap sample with replacement. Under extreme imbalance (100:1), each bootstrap contains approximately 394 majority samples and only 250 minority samples (63% due to sampling with replacement). This introduces high variance in which minority instances each tree observes, leading to divergent tree structures.

TreeSHAP computes Shapley values by averaging contributions across all trees. When tree structures vary substantially due to different minority sample exposure, the averaging process amplifies rather than reduces variance. Small changes in which minority instances are sampled cascade through the ensemble, producing different feature importance rankings.

2) Feature Selection at Splits: Random Forest randomly selects $\sqrt{d}$ features at each split (where $d = 29$ in our case). Under imbalance, splits primarily optimize majority class purity, with minority samples providing weak signal. Different random feature subsets lead to different splitting decisions, creating structural diversity that manifests as explanation instability.

3) Path Dependency in TreeSHAP: TreeSHAP computes Shapley values by analyzing all paths from root to leaf. In Random Forest, the same feature may appear at different tree levels or not at all across trees. When trees have diverse structures, the path-based computation produces feature contributions that depend heavily on tree-specific topology rather than consistent global patterns.

B. Why is XGBoost Stable Despite Being an Ensemble?

XGBoost maintains perfect SHAP stability despite also being a tree ensemble. This stark difference from Random Forest stems from fundamental algorithmic distinctions:

- **Sequential vs. Parallel:** XGBoost builds trees sequentially, each correcting residuals from previous trees. This creates a deterministic optimization trajectory. Random Forest builds trees independently in parallel, introducing structural independence.
- **Gradient-Based Optimization:** XGBoost uses gradient descent, producing consistent trees given fixed data and hyperparameters. Random Forest uses greedy splitting with random feature selection, introducing stochasticity.
- **Weighted Samples:** XGBoost applies sample weights (scale_pos_weight) uniformly across all trees. Random Forest's bootstrap sampling creates variable minority representation per tree.

C. Why is LIME Universally Stable?

LIME's consistent stability across all models stems from its local approximation approach. Unlike SHAP, which analyzes global model structure (tree paths, feature coalitions), LIME:

- Samples perturbed instances around the query point
- Fits a linear surrogate model locally
- Extracts coefficients as feature importance

This local linearization smooths over model complexity and stochasticity. Even when Random Forest produces variable predictions across trees, LIME's linear approximation averages these variations, producing consistent feature importance rankings.

D. The Performance Paradox

Our results show that model performance improves slightly as imbalance increases. This counterintuitive finding contradicts conventional wisdom that imbalance degrades performance.

The explanation lies in our experimental design: we maintain constant minority class size (394 samples) while increasing majority samples. This provides two benefits:

- **Better boundary definition:** More majority samples allow the model to learn a more refined representation of the normal class distribution. Decision boundaries become sharper and more accurate.
- **Consistent minority representation:** The minority class maintains constant representation (394 samples), ensuring

the model has sufficient positive examples regardless of ratio.

This finding has practical implications: when possible, practitioners should maintain adequate minority sample size while allowing majority samples to increase, rather than downsampling excessively.

E. Practical Recommendations

Based on our findings, we provide the following guidance for practitioners:

1) **For Logistic Regression:** Use either SHAP or LIME both provide perfect stability. LIME is faster (0.87s vs 2.34s per instance).
2) **For Random Forest:** Avoid SHAP under severe imbalance (ESS¡0.6 at 50:1+). Use LIME for reliable explanations. Alternatively, consider XGBoost.
3) **For XGBoost:** Both SHAP and LIME work well. SHAP is faster (0.18s vs 0.89s) and offers perfect stability.
4) **General principle:** Always validate explanation stability by generating explanations multiple times with different seeds before deploying to production.

VI. Limitations

This study uses:

- One dataset
- Binary classification
- Two XAI methods

Results may differ in other domains.

VII. Conclusion

This study provides the first systematic investigation of explainable AI stability under class imbalance. Through controlled experiments across five imbalance ratios and three model architectures, we reveal critical stability differences between SHAP and LIME.

Our key findings are:

- SHAP exhibits perfect stability for Logistic Regression and XGBoost but severe instability for Random Forest (ESS=0.538 at 100:1 imbalance)
- LIME maintains perfect stability across all models and imbalance levels
- Explanation drift increases monotonically with imbalance severity
- Tree-based models substantially outperform linear models under extreme imbalance
- Model performance can improve with increasing imbalance when minority class size remains constant

These findings have immediate practical implications. Practitioners deploying Random Forest on imbalanced data should avoid SHAP or validate stability extensively before production deployment. LIME offers a robust alternative with consistent behavior across architectures, albeit with computational tradeoffs.

We hope this work spurs further research into robust explainable AI methods that maintain reliability under the challenging conditions prevalent in real-world deployments.

References

[1] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[2] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[3] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[4] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3319–3328.

[5] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[6] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," *arXiv preprint arXiv:2202.01602*, 2022.

[7] J. Li, L. Xu, L. Yao, and D. Zhang, "Interpretable fraud detection for financial data with a machine learning approach," in *2021 IEEE International Conference on Big Data*, 2021, pp. 2555–2564.

[8] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1379, 2020.

[9] X. Wang, Y. Wang, W. Hsu, and Y. Cai, "Explainable AI for malware detection: A deep learning approach," in *2020 IEEE International Conference on Big Data*, 2020, pp. 3119–3126.

[10] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[11] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.

[12] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *2015 IEEE Symposium Series on Computational Intelligence*, 2015, pp. 159–166.