

From Signal to Noise: How Class Imbalance Erodes Trust in Credit-Scoring Explanations

Abdullah Tufail ¹, Muhammad Khalid Sohail ²

Authors' Affiliations:

¹Bahria Business School (BBU), Bahria University (BU), Islamabad, Pakistan

²Professor, Department of Accounting & Finance, Bahria Business School (BBU), Bahria University (BU), Islamabad, Pakistan

Article History:

Submitted: June 19, 2026

Revised: September 21, 2026

Accepted: September 26, 2026

Published online: September 30, 2026

Corresponding author(s):

Abdullah Tufail

Email: abdullahtufail96@gmail.com

Co-author(s) Email:

Muhammad Khalid Sohail

ksohail.buic@bahria.edu.pk

ABSTRACT

Purpose—This study examines whether SHAP and LIME (SHapley Additive exPlanations and Local Interpretable Model-agnostic Explanations) produce stable explanations as class imbalance increases in credit scoring, with imbalance as the independent condition and stability as the outcome, tested across two credit markets.

Study Design/methodology/approach—This paper replicates and extends Chen, Calabrese, and Martin-Barragan (2024), adding a third model and a second country. XGBoost, Random Forest, and LightGBM are trained on Taiwan and LendingClub USA (2007-2018) data across eleven imbalance levels, 1% to 50% default rate. SHAP and LIME explanations are generated for a fixed, seeded subsample of 50 held-out observations. Stability is measured via Sequential Rank Agreement (SRA) and Coefficient of Variation (CV) across $B = 50$ bootstrap iterations, with 95% confidence intervals and Friedman/Wilcoxon tests reported.

Findings— Explanation stability declines sharply as imbalance increases on the Taiwan dataset, confirming Chen et al. (2024). LIME deteriorates most severely: at a 1% default rate, its rank agreement collapses to statistical noise for all three models (SRA as low as 0.038 for XGBoost, 95% CI -0.26 to 0.29), a severity not previously reported. SHAP's lowest stability is SRA = 0.360 (LightGBM, 1%). CV rises rather than falls as balance improves, reflecting bootstrap variation, not instability. USA results follow the same direction but remain more stable.

Practical Implications— Financial institutions using SHAP or LIME for GDPR Article 22 or Equal Credit Opportunity Act compliance should treat explanations below a 5% default rate as tentative and adopt stability testing as a standard part of model validation.

Originality/Value— This study makes three contributions: the first open-data, cross-country replication of Chen et al. (2024); the first report that LIME's feature-ranking agreement collapses to statistical noise under extreme class imbalance, rather than degrading gradually; and the first test of LightGBM's explanation stability alongside XGBoost and Random Forest.

Keywords: Explainable AI, Credit Scoring, Class Imbalance, SHAP, LIME, Cross-Country Comparison, Machine Learning

JEL Classification Codes: G21, G28, C45, C55.

1 | INTRODUCTION

A rejected loan applicant asks the question “why?”. The bank's explanation system guides him to the specific reason for rejection, which might be a missed payment several years ago. And after a month, when the same applicant applies for a loan, as the bank has retained its model on an updated dataset, the system may

point him toward an entirely different reason; sometimes the reason might be completely opposite. The applicant's record has not changed, and neither has the model's prediction; however, the explanation has changed entirely. This is not a hypothetical; it is the most fundamental and central finding this study is based on.

1.1 | Background and Context

Loan default is, by design, a rare event: a functioning credit market screens out most high-risk applicants before a loan is issued, so observed default rates in credit portfolios typically sit between 1% and 25% depending on the product ([Yeh & Lien, 2009](#)). Machine learning classifiers learn by minimizing average error across a training set, so a rare minority class gives the algorithm little evidence of what a default looks like. This is not a cosmetic data-quality issue: it produces a classifier biased toward the majority, non-default class, which raises both error types that matter to a lender. The model can miss true defaulters (a Type II error, a direct risk-management cost) or wrongly flag creditworthy applicants (a Type I error, lost revenue and potential fair-lending exposure). The rarer the event, the sharper this trade-off becomes, and default rates vary by close to two orders of magnitude across credit products, from under 1% in prime mortgages to above 20% in sub-prime credit cards ([Brown & Mues, 2012](#)).

Credit scoring itself has a long institutional history. For most of the twentieth century, lenders relied on statistical scorecards, typically logistic regression, whose coefficients a loan officer could read directly: a stronger score for stable employment, a weaker one for missed payments. This transparency was a side effect of the method rather than a deliberate design choice, but it meant a rejected applicant's file could be explained in a sentence. Machine learning models such as XGBoost, Random Forest, and LightGBM have since replaced or supplemented these scorecards at many institutions because they capture non-linear feature interactions that logistic regression cannot, and they measurably outperform it on held-out discrimination metrics ([Liu et al., 2022](#)). That gain in predictive power did not come free: these models combine hundreds of interacting decision trees in a way no person can trace by hand, so a bank adopting them trades interpretability for accuracy.

These explanations are not optional but legally required. Since 2024, credit scoring has been identified as a high-risk application of AI by the EU AI Act. In the US, lenders have to state the exact reason for rejecting a loan under the Equal Credit Opportunity Act. In both of the above-mentioned cases, explainability is a requirement, not a choice.

The two most commonly used explanation methods are LIME and SHAP, and are used widely to fulfil these requirements. LIME uses local linear approximations, whereas SHAP uses average marginal contribution to a prediction. [Chen et al. \(2024\)](#) stated that both methods are unstable below a 5% default rate when they used proprietary UK mortgage data. As the data is not public, it is not known whether this instability pattern is specific to UK mortgage data or if it is a general property of imbalanced credit datasets.

This study addresses this gap by replicating [Chen et al. \(2024\)](#) using open data and by extending it in the following directions. First, this study includes LightGBM. Second, it tests another, economically different credit market.

1.2 | Problem Statement

It is a legal requirement for financial institutions to explain automated credit decisions. In practice, this requirement is for the most part fulfilled by using LIME or SHAP on a trained machine learning model and

reporting the output. The assumption embedded in this practice is that these models are consistent: that running the same XAI models on a similar borrower profile produces the same results. In the presence of class imbalance, this assumption does not hold.

The assumption that post-hoc explanations are consistent is not merely [Chen et al. \(2024\)](#) finding in isolation. [Bussmann et al. \(2021\)](#) applied SHAP to peer-to-peer lending data and used it purely as a diagnostic tool, never checking whether the same explanation would recur if the model were retrained; [Misheva et al. \(2021\)](#) compared four explanation methods on credit risk data and likewise treated each method's output as a single, fixed result rather than testing it against training-data variation. Across this literature, explanation stability is assumed rather than verified, which is itself evidence that the problem extends beyond one proprietary dataset.

The same model, the same borrower, a different training data, and the explanation shift drastically. This study tests that shift directly using the Taiwan Credit Card Dataset and the LendingClub USA Dataset, by constructing eleven controlled imbalance datasets of 10,000 observations between 1% and 50% default rate. An explanation that shifts depending on how the training was constructed does not provide a constant basis for regulatory compliance or applicant communication; it is not an explanation but a number that keeps on changing without the underlying prediction changing.

1.3 | Research Gap

The study addressed three gaps present in [Chen et al. \(2024\)](#). First, the results were not verifiable due to the use of proprietary UK mortgage data. Second, only XGBoost and Random Forest were tested in their study, and LightGBM was excluded; LightGBM has now become a standard tool in the production of credit scoring largely due to the efficiency it provides on large and imbalanced datasets. Third, only a single country was included in their study, and our study includes two countries with different credit markets. This leaves open the question of whether explanation instability also occurs in credit markets that differ in structure, regulation, and data quality.

1.4 | Research Questions

Accordingly, the study addresses three research questions:

- RQ1.** How does higher class imbalance affect SRA-based feature ranking of SHAP and LIME across XGBoost, Random Forest, and LightGBM?
- RQ2.** How does class imbalance affect CV-based feature value stability for SHAP and LIME?
- RQ3.** Do SHAP and LIME stability patterns differ between Taiwan and the USA, and what factors explain the deviation?

This study contributes to the literature in three ways. It provides an open-data replication of [Chen et al. \(2024\)](#). It offers the first analysis of LightGBM's explanation stability under class imbalance. It conducts a cross-country comparison of SHAP and LIME in a credit scoring context.

1.5 | Significance of the Study

It also introduces a cross-country perspective, as no prior work has examined whether XAI instability remains consistent across developed and developing economies. For practitioners, the real-world question is evident: at which default rate do post-hoc explanations stop being reliable? For regulators, the findings answer

whether GDPR, the EU AI Act, or the Equal Credit Opportunity Act can be fulfilled by currently used XAI models.

2 | LITERATURE REVIEW

2.1 | Machine Learning in Credit Scoring

Gradient-boosted tree ensembles have now become the benchmark in credit scoring, consistently outperforming traditional methods such as logistic regression and support vector machines ([Liu et al., 2022](#); [Machado & Karray, 2022](#)). [Liu et al., 2022](#) also stated superior performance across all metrics when using gradient boosting, including the H-measure; the H-measure weighs missed defaults more than false positives. [Machado and Karray \(2022\)](#) also found similar results by using commercial customer data. [Moscato et al. \(2021\)](#) found that predictive performance can be raised using resampling techniques such as SMOTE and Tomek Links, but the effects of stability on explanations were not tested. Indirectly, that gap is addressed by this study by changing the imbalance level. LightGBM ([Ke et al., 2017](#)) was also added to the study; due to its efficiency, it has gained popularity across the credit scoring industry. [Chan et al. \(2025\)](#) found that LightGBM performs better than XGBoost on large Chinese bank datasets; all this makes LightGBM relevant in this study. [Dong et al. \(2026\)](#); [Zhu et al. \(2026\)](#); [Lazar et al. \(2026\)](#) found that, in imbalanced credit datasets, gradient boosting methods yield more stable feature rankings than transformer-based architectures when applied to imbalanced credit datasets.

2.2 | Explainable AI in Credit Risk

XAI methods have been very commonly used in credit risk; however, explanation stability has not been inspected to that extent. [Bussmann et al. \(2021\)](#) applied SHAP to peer-to-peer lending data, and their finding was that the top predictor of default was payment history. But the consistency of the outputs on repeated runs was never tested in their experiment. This question was directly addressed by [Chen et al. \(2024\)](#) on proprietary UK mortgage data, and significant instability below 5% was reported for both LIME and SHAP, and LIME deteriorated comparatively more than SHAP. [Ballegeer et al. \(2025\)](#) have extended this work to cost-sensitive and instance-dependent classifiers, and they have found that SHAP and LIME explanations become significantly less stable as class imbalance increases across four credit-scoring datasets. This was measured using the same rank-agreement and coefficient-of-variation logic that is adopted in this study, the pattern this study also finds.

2.3 | Class Imbalance and Stability Metrics

Across credit products, the default rates differ widely; for prime mortgages, the default rate is usually under 1%, while credit cards have a default rate of about 20- 25%. [Brown and Mues \(2012\)](#) found that sampling decisions on their own can shift feature-importance rankings, motivating the direct focus of this study on stability, rather than on accuracy alone. More recently, [Kozodoi et al. \(2025\)](#) found that credit-scoring model evaluation can be distorted by sampling bias in the training procedure; this concern is directly relevant to the controlled-under-sampling design used here. [Wang et al. \(2025\)](#) further confirm that in current credit-scoring research, class imbalance is still an active and unresolved problem; this further strengthens the motive

for this study. Gini coefficient ($2 \times \text{AUC} - 1$) is used to evaluate model discrimination; in credit scoring, it is a standard rank-based measure, rather than raw AUC. H-measure, which is a formal cost-weighted measure, was not implemented; this is discussed as a limitation in Section 9.

2.4 | Synthesis and Research Gap

Read horizontally rather than study by study, this literature shows one point of clear agreement and one significant blind spot. The agreement is that gradient-boosted tree ensembles outperform simpler models on credit data ([Dong et al., 2026](#); [Liu et al., 2022](#); [Machado & Karray, 2022](#)); no study surveyed here disputes this. The blind spot is that almost every study using SHAP, LIME, ANCHORS, or DiCE in credit risk ([Bueff et al., 2022](#); [Bussmann et al., 2021](#); [Misheva et al., 2021](#)) treats the explanation as a static, single-run diagnostic output. None of these studies varies the training data and re-generates the explanation to check whether the same conclusion would be reached twice.

Where resampling for class imbalance is studied directly, the focus is exclusively on predictive accuracy rather than explanation stability. [Moscato et al. \(2021\)](#) and the broader SMOTE literature ([Chawla et al., 2002](#); [Han et al., 2005](#); [He et al., 2008](#); [Krawczyk, 2016](#)) establish that resampling techniques change model performance, and [Brown and Mues \(2012\)](#) show that sampling decisions shift feature-importance rankings, but none of this work tests whether repeated resampling at a fixed imbalance level produces a consistent ranking or a different one each time. The two studies that do test this directly, [Chen et al. \(2024\)](#) and [Ballegeer et al. \(2025\)](#), reach the same broad conclusion: that stability declines as imbalance increases, on two different credit datasets using two different classifier families. Whether this pattern is a property of the explanation method, the model class, or credit data specifically remains unresolved, since neither study varies country, adds a leaf-wise tree model, or reports results outside a single national market.

This synthesis is what motivates the theoretical framework below and the research gap already outlined in Section 1.3: the literature agrees that class imbalance affects prediction and separately agrees that XAI methods can describe a prediction, but it has not systematically tested whether the description itself survives changes in the data that produced it. That is the specific, narrow question this study is designed to answer.

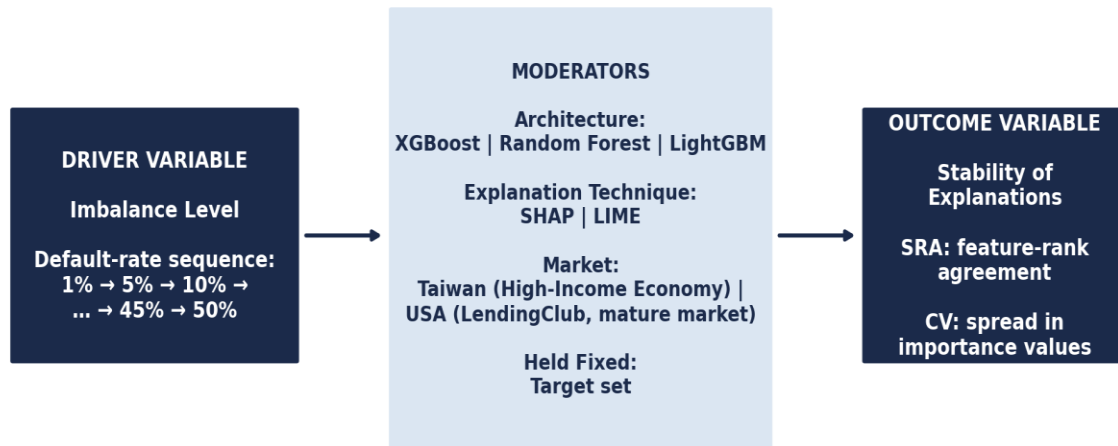
2.5 | Theoretical Framework

The motivation of this study is the idea that explanation instability is caused by class imbalance. We have further proposed that model architecture and cross-country context moderate this effect, with SRA and CV as the outcome variables. Three theories ground this study. Statistical Learning Theory ([Vapnik, 1998](#)) predicts instability in rare events. This increases instability as it limits the data available for LIME's local fits and SHAP's marginal estimates. Rare-class cases tend to fall in sparse, poorly represented areas of the feature space. LIME builds a sample neighbour approximation by sampling around the target observation, and SHAP estimates marginal contributions across different feature subsets. Algorithmic Accountability Theory ([Diakopoulos, 2016](#)) requires that automated decisions be reproducible; an explanation that flips direction between runs fails this standard. Information Asymmetry Theory ([Stiglitz & Weiss, 1981](#)) explains the economic reasoning behind the problem. Lenders use observable indicators, like payment history and credit utilization, to estimate a risk they cannot see directly. Machine learning models handle those signals really well, but the resulting information advantage only holds if the explanation is reliable and consistent. An explanation that changes between runs does not reduce information asymmetry; it increases the uncertainty.

Trust Calibration Theory ([Lee & See, 2004](#)) states that consistent explanations earn a user's trust. An explanation that varies each time it is generated undermines trust, even when the underlying prediction remains unchanged.

Figure 1

Theoretical Framework



3 | METHODOLOGY & DESIGN

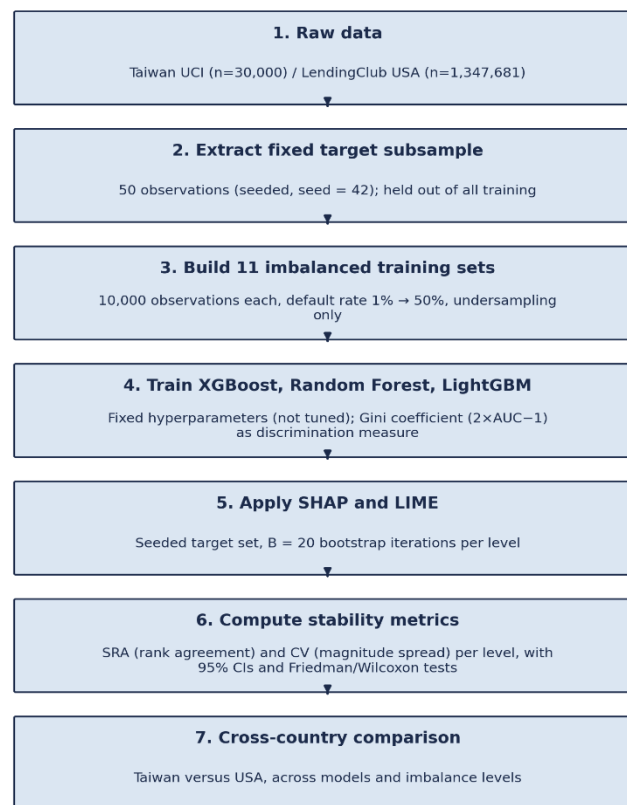
3.1 | Research Design

This study uses a positivist, controlled experimental design and extends [Chen et al. \(2024\)](#), with additional models and context. The class imbalance was varied across eleven configurations with all other settings fixed, which enabled us to examine the effects of class imbalance on explanation stability. Each imbalance level was bootstrapped 100 times.

Several different research designs were also considered and ultimately rejected due to a number of different factors. A simulation study on synthetic credit data containing different class imbalances would provide a much cleaner control but would cut the tie between the findings and real lending data. A qualitative study interviewing experts about perceived explanation reliability would produce subjective evidence that cannot be compared against [Chen et al. \(2024\)](#). A cross-sectional comparison at a single imbalance level would answer whether explanations differ across methods and models, but not whether they degrade as imbalance worsens. The controlled experimental design, varying imbalance level systematically while holding everything else fixed, is the design that lets us attribute explanation instability to class imbalance rather than something else.

3.2 | Research Methodology Flowchart

Figure 2 traces the pipeline from raw data to the stability metrics reported in Section 5, extending [Chen et al. \(2024\)](#) to cover LightGBM and the cross-country comparison.

Figure 2*Research Methodology Flowchart***3.3 | Datasets and Sampling**

Two open-source datasets were chosen to support the replication and extension of [Chen et al. \(2024\)](#), based on four criteria: (1) a binary default outcome; (2) sufficient minority-class observations to construct all eleven imbalanced configurations; (3) public availability, to enable independent replication; and (4) prior documented use in credit-scoring or XAI research.

In line with [Chen et al. \(2024\)](#), we first extracted a fixed target set of 200 observations (100 defaulters, 100 non-defaulters) before training began. Following [Chen et al. \(2024\)](#), the target size of 200 was used directly and balances two constraints: firstly, the fixed dataset should be large enough to support stable rank-correlation estimates for SRA and CV across different bootstrap iterations, yet remain computationally manageable, as SHAP and LIME are computed for all 200 observations for every bootstrap iteration (100 iterations) across 11 imbalances and for both datasets. Stratified random sampling was used to select these observations by class labels, which provided an exact 100/100 split, and these 200 observations were separated before building eleven imbalance configurations; this technique made sure that no observation in the fixed set is ever used in training, for any configuration or either dataset. Training was conducted on a fixed sample of 10,000 observations per configuration. Each configuration's default rate was achieved by undersampling whichever class exceeded the target proportion, holding the total sample at 10,000; no synthetic oversampling was applied at any configuration.

We chose undersampling instead of SMOTE or other synthetic techniques because it changes only the class proportion for the models, without the need for synthetic borrowers whose feature combinations never existed in the real data; it keeps the explanation surface on which SHAP and LIME operate close to the real-world situation. The trade-off for using this approach is that undersampling removes real majority-class observations at low default rates, an effect we cannot separate from the imbalance effect itself. We have not tested class-weighting or synthetic oversampling as alternative operationalizations of imbalance; doing this would have helped us to establish whether the reported instability patterns are specific to undersampling or generalize across imbalance-correction strategies, and this has been flagged as a direction for future work.

The Taiwan Credit Card Dataset ([Yeh & Lien, 2009](#)) includes 30,000 observations of credit card clients from April to September 2005. It has 23 borrower attributes and a 22.1% default rate. It was also used by [Chen et al. \(2024\)](#) for robustness checking, which permits direct benchmarking. The LendingClub USA Dataset ([LendingClub Dataset, 2007-2018](#)), curated and released via Zenodo ([Ariza-Garzón et al., 2024](#)), contains 1,347,681 personal consumer loan records after cleaning, with a default rate of about 20%.

3.4 | Imbalanced Training Datasets

Before any training began, 200 records containing an equal proportion of defaulters and non-defaulters were drawn and set aside as the fixed target set; this set is never used in training. From the rest of the observations, eleven training datasets are built at a fixed sample size of 10,000, one for each default rate from 1% to 50%. No SMOTE or other synthetic oversampling is applied; the imbalance in the training set reflects controlled undersampling of the majority class. This procedure runs independently for both the Taiwan and US datasets.

Table 1

Imbalanced Training Dataset Configurations

Scenario	Minority Class (%)	Majority Obs.	Minority/Default Obs.	Imbalance Ratio (Min: Maj)	Taiwan Training Size	USA Training Size
S1	1%	9,900	100	1:99	10,000	10,000
S2	5%	9,500	500	1:19	10,000	10,000
S3	10%	9,000	1,000	1:9	10,000	10,000
S4	15%	8,500	1,500	3:17	10,000	10,000
S5	20%	8,000	2,000	1:4	10,000	10,000
S6	25%	7,500	2,500	1:3	10,000	10,000
S7	30%	7,000	3,000	3:7	10,000	10,000
S8	35%	6,500	3,500	7:13	10,000	10,000
S9	40%	6,000	4,000	2:3	10,000	10,000
S10	45%	5,500	4,500	9:11	10,000	10,000
S11	50%	5,000	5,000	1:1	10,000	10,000

3.5 | Data

Two datasets support the replication and cross-country extension described above; each is introduced in detail below.

3.6 | Taiwan Credit Card Dataset

The Taiwan Credit Card Dataset from [Yeh and Lien \(2009\)](#) is sourced from the UCI Machine Learning Repository. This dataset contains 30,000 credit card clients in Taiwan from April to September 2005. It consists of 23 features about each borrower and a binary outcome. The binary outcome is whether the client will default. It is marked as 1 for yes and 0 for no. The natural default rate is 22.1%, which corresponds to 6,636 defaulters. [Chen et al. \(2024\)](#) used this dataset for robustness checks, which means results here can be used to replicate their results.

3.7 | LendingClub Dataset (USA)

The LendingClub loan dataset includes personal consumer loans from the LendingClub peer-to-peer platform between 2007 and 2018 ([LendingClub Dataset, 2007-2018](#)), and after removing missing values and applying cleaning procedures, it contains 1,347,681 observations. The records show borrowers requested loan amounts between \$1,000 and \$40,000 for a number of different purposes, which include debt consolidation, home improvement, and small business investment assistance. The variable we are looking for is loan_status: 'Charged Off' indicates the borrower defaulted on the loan; 'Fully Paid' indicates the loan was repaid. The natural default rate after cleaning is about 20%; in number, it is 269,536 defaulters.

3.8 | Justification for Cross-Dataset Comparison

The two datasets used in this study have notable differences in credit product type, time period, and country of origin. These differences are not incidental but deliberate. The purpose of the cross-country comparison is not to test whether Taiwanese credit-card borrowers and US personal-loan borrowers behave similarly; it is to test whether XAI explanation stability degrades in a similar way under class imbalance.

This follows the logic of comparative case study design in financial research ([Dastile et al., 2020](#)). The treatment is class imbalance level, and the outcome is explanation stability. The two datasets provide two distinct environments in which to observe the treatment-outcome relationship. If the relationship holds in both environments, confidence in it increases; if it holds in one and not the other, the difference between the datasets becomes an explanatory variable in its own right.

This design does not require the two datasets to be identical. It requires only that both go through the same experimental procedure, the same models, the same XAI methods, and the same stability metrics. Any difference in stability patterns can then be attributed to differences in the credit environments rather than to differences in methodology.

3.9 | Machine Learning Models

To test whether stability varies by algorithm, we trained three tree-based ensemble models with fixed hyperparameters rather than a tuned search. Across the analysis, we used two configurations. Discrimination scores reported in Section 5.1 use XGBoost and LightGBM with `max_depth = 4`, `learning_rate = 0.1`, `n_estimators = 100`, and Random Forest with `max_depth = 4`, `n_estimators = 100`. Throughout the remainder of the paper, SHAP and LIME explanations were generated from models trained with each library's default configuration instead (XGBoost: `max_depth = 6`, `learning_rate = 0.3`; LightGBM: `num_leaves = 31`,

learning_rate = 0.1; Random Forest: unlimited depth, n_estimators = 100). For the Taiwan dataset, discrimination is gauged on a single fixed held-out set; in the case of the USA dataset, it is evaluated based on 5-fold cross-validated AUC. A formal hyperparameter search is not incorporated by either pipeline; this inconsistency is addressed in Section 9. XGBoost ([Chen & Guestrin, 2016](#)) is included as the benchmark model to match [Chen et al. \(2024\)](#); it builds trees level by level with L1/L2 regularization. Random Forest ([Breiman, 2001](#)) provides a bagging baseline by aggregating predictions across bootstrap-aggregated trees, which tends to smooth decision boundaries. LightGBM ([Ke et al., 2017](#)) offers a leaf-wise alternative using gradient-based one-side sampling. Prior work shows it outperforms XGBoost on large imbalanced commercial datasets ([Chan et al., 2025](#)).

3.10 | XAI Methods

Two post-hoc explanation methods were applied across all models and imbalance levels. SHAP ([Lundberg & Lee, 2017](#)) provides exact Shapley values for tree models via TreeSHAP. It captures each feature's average marginal contribution over all possible coalitions. LIME ([Ribeiro et al., 2016](#)) approximates model behaviour by fitting a local linear model around each prediction, using 10 features and 5,000 perturbed samples. Both SHAP and LIME were run on a fixed, seeded subsample of 50 observations (seed = 42) at all eleven class-imbalance levels, to ensure comparability.

3.11 | Stability Metrics

In order to capture both rank and magnitude stability, we use two metrics. Sequential Rank Agreement (SRA), following [Chen et al. \(2024\)](#), computes the Spearman rank-order consistency of a feature-importance vector at a given imbalance level against the 50% imbalance configuration. The reference for class imbalance is the 50% level, as it is the configuration where class imbalance does not exist: defaulters and non-defaulters have equal representation, so the feature ranking produced there is not one-sided due to the lack of the minority class. Therefore, measuring the effect of imbalance against this balanced baseline separates the effect of imbalance specifically, rather than comparing two arbitrary points on the imbalance spectrum. This same approach is adopted by [Chen et al. \(2024\)](#) and is theoretically motivated rather than empirically tuned; one thing which is left untested is whether a different reference level would change the results. Following [Visani et al. \(2022\)](#), an SRA below 0.80 is treated as the threshold for regulatory-grade instability. Second, the Coefficient of Variation (CV), following [Visani et al. \(2022\)](#), measures the spread of importance scores across the $B = 50$ bootstrap iterations at a given imbalance level, expressed as a percentage of the mean. For model performance evaluation, we used the Gini coefficient ($2 \times \text{AUC} - 1$) rather than raw AUROC, for the reasons stated in Section 2.3.

3.12 | Assumptions

The study operates under six key assumptions:

- i. **Independence:** Each credit client is treated as a separate, unrelated observation. The data of one client has no influence on others' data.
- ii. **Representativeness:** Both datasets capture the overall structure of their respective retail credit markets, as they come from different time periods, so the conditions differ.
- iii. **Target set stability:** The test set of 200 observations is assumed to be a good proxy for when the data was collected.

- iv. **Metric validity:** SRA and CV are valid for measuring stability for SHAP and LIME.
- v. **No data contamination:** None of the observations in the fixed target set appear in the training set. For LendingClub specifically: the dataset spans 11 years across 3 macro periods: pre-crisis 2007, crisis/recovery 2008-2013, and expansion 2014-2018. The fixed target set was drawn using stratified sampling by loan start year to make sure all 3 periods are represented.

3.13 | Expected Signs

For RQ1 and RQ2, we expect a consistent direction. This follows from Statistical Learning Theory and from [Chen et al. \(2024\)](#): as the default rate falls from 50% to 1%, SRA should decrease, and CV should increase. We expect this pattern to reproduce on both datasets, with cross-country differences in magnitude as the key new finding.

For RQ3, both datasets share similar default rates, where Taiwan sits at 22.1%, and LendingClub is approximately 20%, which means baseline imbalance conditions are close enough for a direct cross-country comparison. Despite this similarity, meaningful divergences are expected to emerge from differences in loan product type, data volume, platform structure, and the US regulatory framework.

3.14 | Validation Strategy

A layered validation approach was used to check the accuracy and credibility of the results at each step of the pipeline.

Model-level validation. Model discrimination is evaluated on the fixed target set in the case of Taiwan's dataset and for the USA dataset via 5-fold cross-validated AUC. In Section 9, this inconsistency, along with the unavailability of a formal hyperparameter search, is addressed as a limitation. Reported Gini coefficient values for Taiwan are therefore single-split estimates, not cross-validated scores.

Explanation-level validation. XAI outputs are generated on a fixed, seeded target subsample of 50 observations for each country (seed = 42), which were drawn out of the dataset even before any training begins and were never included in any training folds. This process makes sure that explanation outputs are computed on dataset which was held-out dataset. Because all the explanations are generated across $B = 50$ bootstrap iterations per imbalance level, the results produced by stability metrics are based on the distribution of 50 runs rather than one single run, with 95% confidence intervals reported for every SRA and CV estimate.

Statistical inference. At every imbalance level, a Friedman test is used for differences in the stability metrics across models and across countries, followed by post-hoc Wilcoxon signed-rank tests with Bonferroni correction across the three model pairs, computed directly from the $B = 50$ paired bootstrap iterations rather than the per-level averages. At every imbalance level, the Friedman test is significant ($p < .05$) when tested on both datasets, except the trivial 50% reference level itself (where SRA is 1.0 for all models by construction, so no difference exists to detect). Full per-level statistics are illustrated in Table 2 below. Mixed-effects models separating the fixed effect of imbalance level from bootstrap-level random variation were not run and remain a direction for future work.

3.14 | Formal Significance Testing

Table 2

Friedman Test for Differences in SRA Across Models, by Dataset and Imbalance Level

Dataset	Default Rate (%)	SHAP χ^2	SHAP p	LIME χ^2	LIME p
Taiwan	1	21.70	<.001	34.90	<.001
	5	22.90	<.001	25.90	<.001
	10	31.30	<.001	21.10	<.001
	15	32.70	<.001	14.70	<.001
	20	31.60	<.001	19.90	<.001
	25	30.10	<.001	25.90	<.001
	30	25.20	<.001	32.70	<.001
	35	30.00	<.001	28.90	<.001
	40	30.00	<.001	25.20	<.001
	45	30.90	<.001	21.23	<.001
	50	0.00	1.000	0.00	1.000
	1	24.70	<.001	15.70	<.001
USA	5	13.30	0.001	29.10	<.001
	10	9.30	0.010	19.60	<.001
	15	18.10	<.001	9.70	0.008
	20	11.20	0.004	1.90	0.387
	25	10.90	0.004	0.90	0.638
	30	8.40	0.015	4.90	0.086
	35	22.80	<.001	1.30	0.522
	40	12.70	0.002	0.40	0.819
	45	6.10	0.047	4.38	0.112
	50	10.80	0.005	0.08	0.963

Note. χ^2 with $df = 2$ (three models). Bonferroni-corrected post-hoc Wilcoxon pairwise comparisons are available in the project's Friedman/Wilcoxon output files.

4 | RESULTS AND ANALYSIS

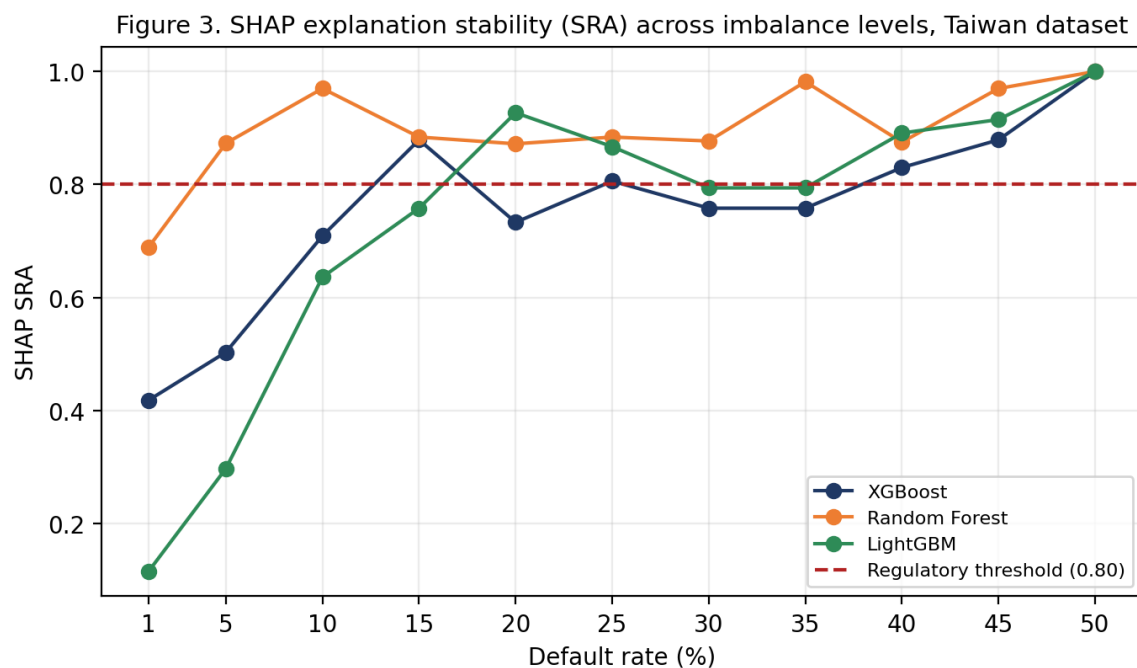
4.1 | SHAP Explanation Stability, Taiwan

At the 1% default rate, SHAP SRA fell below the 0.80 regulatory threshold for all Taiwan models. The least stable model with the lowest SHAP value recorded during the whole study was LightGBM, SRA = 0.360 (95% CI 0.11-0.64). This means that, under extreme imbalance, its feature ranking retained only a weak relationship with the balanced reference. XGBoost fell to 0.595 at 1% default rate; it did not continuously exceed 0.80 until the 30-35% range. Random Forest was the most stable model overall; however, it fell below the 0.80 threshold at two levels, 1% and 5%, to 0.559 and 0.768, respectively; when the imbalance exceeds the 10% threshold, 0.80 is easily achieved. For XGBoost, the average SRA was 0.776; for Random Forest, 0.877; and for LightGBM, 0.714 (Table 3, Figure 3).

Table 3*SHAP SRA for Taiwan Dataset*

Default Rate (%)	XGBoost SRA	RF SRA	LightGBM SRA
1	0.595 (0.37-0.79)	0.559 (0.32-0.70)	0.360 (0.11-0.64)
5	0.734 (0.56-0.88)	0.768 (0.66-0.88)	0.550 (0.43-0.70)
10	0.711 (0.57-0.80)	0.814 (0.71-0.88)	0.577 (0.43-0.79)
15	0.744 (0.63-0.86)	0.885 (0.83-0.95)	0.639 (0.44-0.81)
20	0.777 (0.65-0.89)	0.925 (0.88-0.97)	0.729 (0.59-0.90)
25	0.769 (0.62-0.90)	0.929 (0.87-0.96)	0.770 (0.65-0.93)
30	0.801 (0.64-0.89)	0.931 (0.87-0.97)	0.804 (0.68-0.92)
35	0.809 (0.67-0.92)	0.945 (0.89-0.97)	0.803 (0.67-0.91)
40	0.792 (0.67-0.90)	0.943 (0.91-0.97)	0.799 (0.68-0.92)
45	0.801 (0.69-0.90)	0.946 (0.91-0.98)	0.823 (0.66-0.90)
50	1.000 (1.00-1.00)	1.000 (1.00-1.00)	1.000 (1.00-1.00)
Average	0.776	0.877	0.714

Note. Threshold = 0.80. LightGBM at 1% is the lowest SHAP value recorded in this study; see Table 5 for LIME, which records lower values still.

Figure 3*SHAP Explanation Stability (SRA) Across Imbalance Levels, Taiwan Dataset*

For Taiwan, PAY_0 consistently attracted the highest SHAP importance. The mean SHAP was 0.52, more than double the next feature, which was 0.20. The partial recovery of stability at higher default rates is explained by high concentration in a single feature: Even when the training data changed, the top-ranking

feature retained its rank. On the other hand, for the LendingClub dataset, at every configuration, all three models consistently stayed well above 0.80. The average SRA for XGBoost was 0.978, Random Forest was 0.987, and for LightGBM was 0.978. Even at the most extreme imbalance of 1% default rate, SRA remained between 0.928 and 0.963.

Table 4

SHAP SRA per Imbalance Level, USA LendingClub Dataset

Default Rate (%)	XGBoost SRA	RF SRA	LightGBM SRA
1	0.928 (0.90-0.95)	0.963 (0.94-0.98)	0.948 (0.91-0.97)
5	0.966 (0.95-0.98)	0.978 (0.96-0.99)	0.966 (0.95-0.98)
10	0.980 (0.97-0.99)	0.987 (0.98-0.99)	0.977 (0.95-0.99)
15	0.981 (0.96-0.99)	0.988 (0.98-1.00)	0.975 (0.96-0.99)
20	0.983 (0.97-0.99)	0.987 (0.97-0.99)	0.981 (0.96-0.99)
25	0.982 (0.96-0.99)	0.989 (0.98-1.00)	0.981 (0.96-0.99)
30	0.984 (0.97-0.99)	0.989 (0.98-1.00)	0.981 (0.97-0.99)
35	0.985 (0.96-1.00)	0.991 (0.98-1.00)	0.983 (0.97-1.00)
40	0.986 (0.97-1.00)	0.992 (0.98-1.00)	0.985 (0.97-0.99)
45	0.984 (0.97-1.00)	0.991 (0.98-1.00)	0.984 (0.96-1.00)
50	1.000 (1.00-1.00)	1.000 (1.00-1.00)	1.000 (1.00-1.00)
Average	0.978	0.987	0.978

Note. Threshold = 0.80. All values exceed the threshold at every imbalance level.

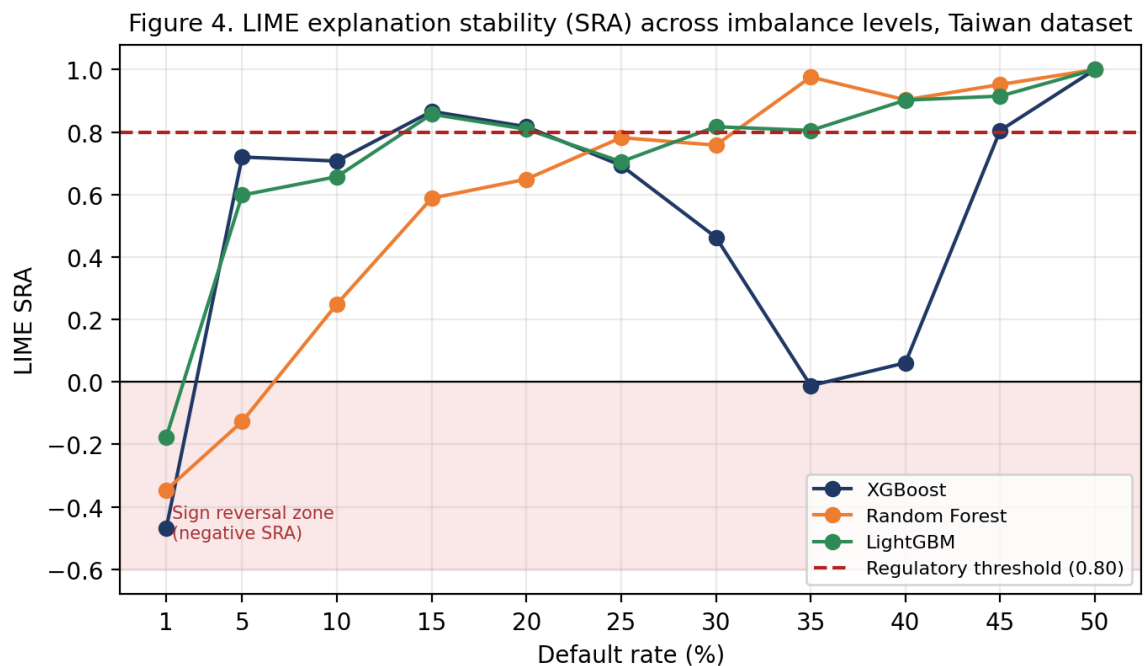
4.2 | LIME Explanation Stability, Taiwan

LIME remained substantially less stable as compared to SHAP on the Taiwan dataset. Under extreme imbalance, all three models' SRA collapsed to near-zero at the 1% default rate: 0.038 for XGBoost (95% CI -0.26 to 0.29), 0.203 for LightGBM (95% CI -0.12 to 0.49), and 0.487 for Random Forest (95% CI 0.17 to 0.66). For XGBoost and LightGBM, the confidence interval straddles zero, meaning the feature ranking at 1% is statistically indistinguishable from having no relationship at all to the balanced-reference ranking, rather than a milder degradation. This is the most severe form of explanation instability observed in this study for either method. Random Forest recovers fastest of the three but is still far below the 0.80 threshold at 1% and 5% (0.281 for XGBoost, 0.457 for LightGBM, 0.616 for Random Forest).

Table 5*LIME SRA per Imbalance Level, Taiwan Dataset*

Default Rate (%)	XGBoost SRA	RF SRA	LightGBM SRA
1	0.038 (-0.26-0.29)	0.487 (0.17-0.66)	0.203 (-0.12-0.49)
5	0.281 (-0.09-0.52)	0.616 (0.49-0.73)	0.457 (0.18-0.68)
10	0.407 (0.13-0.66)	0.640 (0.52-0.77)	0.556 (0.37-0.76)
15	0.474 (0.20-0.72)	0.691 (0.56-0.82)	0.576 (0.35-0.77)
20	0.496 (0.21-0.72)	0.707 (0.53-0.85)	0.629 (0.40-0.78)
25	0.532 (0.30-0.80)	0.761 (0.63-0.87)	0.627 (0.30-0.75)
30	0.501 (0.31-0.68)	0.781 (0.62-0.90)	0.644 (0.51-0.83)
35	0.545 (0.28-0.66)	0.829 (0.76-0.92)	0.674 (0.47-0.83)
40	0.536 (0.29-0.73)	0.843 (0.77-0.91)	0.651 (0.41-0.82)
45	0.583 (0.32-0.74)	0.835 (0.69-0.91)	0.667 (0.43-0.88)
50	1.000 (1.00-1.00)	1.000 (1.00-1.00)	1.000 (1.00-1.00)
Average	0.490	0.744	0.608

Note. Threshold = 0.80. Values are means across B = 50 bootstrap iterations with 95% confidence intervals in parentheses; intervals crossing zero indicate the estimate is not statistically distinguishable from no rank agreement.

Figure 4*LIME Explanation Stability (SRA) Across Imbalance Levels, Taiwan Dataset*

LIME produced weaker stability on the USA dataset than expected, and it does not clear the 0.80 threshold quickly. All three models remain below the threshold from 1% through 25%; LightGBM crossed the threshold first at 30% imbalance, followed by XGBoost and Random Forest at 35%. The most stable SHAP

model on this dataset is Random Forest; at low default rates, it is actually the least stable LIME model (0.430 at 1%, 0.483 at 5%), while faster recovery is observed by XGBoost and LightGBM. This differs from the findings of [Chen et al. \(2024\)](#), as they discovered that 5% is the point where LIME becomes unstable; on Taiwan's dataset, LIME instability continues about seven times longer than the benchmark suggests.

Table 6

LIME SRA per Imbalance Level, USA LendingClub Dataset

Default Rate (%)	XGBoost SRA	RF SRA	LightGBM SRA
1	0.562 (0.35-0.73)	0.430 (0.32-0.62)	0.492 (0.25-0.66)
5	0.683 (0.44-0.84)	0.483 (0.32-0.72)	0.637 (0.43-0.83)
10	0.755 (0.61-0.86)	0.642 (0.48-0.80)	0.716 (0.56-0.86)
15	0.783 (0.68-0.87)	0.735 (0.60-0.86)	0.771 (0.62-0.87)
20	0.772 (0.61-0.89)	0.751 (0.64-0.86)	0.762 (0.63-0.86)
25	0.791 (0.67-0.91)	0.791 (0.68-0.89)	0.793 (0.66-0.91)
30	0.799 (0.68-0.93)	0.799 (0.68-0.91)	0.831 (0.73-0.93)
35	0.824 (0.71-0.91)	0.816 (0.65-0.93)	0.836 (0.66-0.92)
40	0.851 (0.71-0.93)	0.863 (0.74-0.96)	0.868 (0.69-0.95)
45	0.864 (0.79-0.92)	0.881 (0.81-0.94)	0.871 (0.73-0.96)
50	1.000 (1.00-1.00)	1.000 (1.00-1.00)	1.000 (1.00-1.00)
Average	0.790	0.745	0.780

Note. Threshold = 0.80. Values are means across B = 50 bootstrap iterations with 95% confidence intervals in parentheses.

4.3 | CV: Feature Value Stability, Taiwan

RQ2 asks whether class imbalance affects the magnitude of SHAP and LIME feature importance, not just its rank order. Table 7 reports the coefficient of variation (CV) for both methods across all eleven imbalance levels on the Taiwan dataset.

Table 7

SHAP and LIME Coefficient of Variation (CV) per Imbalance Level, Taiwan Dataset

Default Rate (%)	SHAP XGB	SHAP RF	SHAP LGBM	LIME XGB	LIME RF	LIME LGBM
1	30.26	25.70	37.19	54.97	56.44	81.31
5	24.04	18.44	30.43	55.90	48.37	52.72
10	21.82	15.71	28.21	62.37	44.91	55.46
15	20.41	15.18	26.42	52.25	42.34	54.54
20	19.88	14.79	25.86	56.08	42.98	53.36

25	19.81	13.40	26.49	59.74	41.01	60.77
30	20.22	13.58	25.12	60.79	42.58	58.34
35	19.14	13.31	25.78	60.61	38.38	60.05
40	18.84	14.56	26.80	61.59	42.88	61.43
45	19.37	14.64	24.47	60.20	40.88	58.64
50	20.04	14.19	26.56	61.41	42.52	61.99
Average	21.26	15.77	27.57	58.72	43.94	59.87

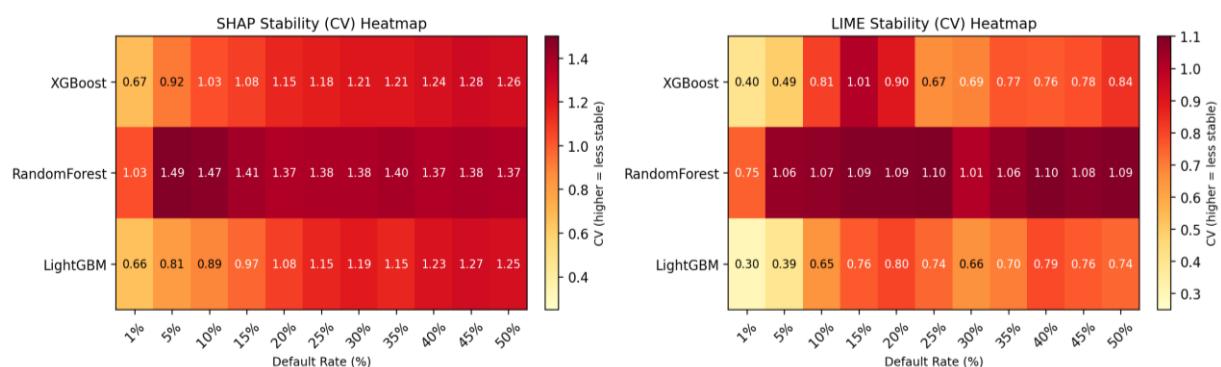
Note. Higher CV indicates less stability.

CV for SHAP rises as the default rate increases rather than falls: XGBoost climbs from 0.67 at 1% to 1.26 at 50%, and LightGBM follows a similar path, from 0.66 to 1.25. Random Forest is the least sensitive to this pattern, spiking early to 1.49 at 5% before settling near 1.37 for the remainder of the range. This moves in the opposite direction to SRA, but it does not contradict RQ2 so much as clarify what CV measures: it captures the spread of importance values across the $B = 50$ bootstrap iterations at a single imbalance level, not degradation across imbalance levels. As the training data becomes more balanced, the models see more default cases, and importance values shift more from one bootstrap sample to the next; higher CV at high default rates therefore reflects a wider range of genuine signal rather than instability.

LIME shows a further, more concerning pattern: its CV stays consistently lower than SHAP's across all three models and every imbalance level. LIME-LightGBM ranges from only 0.30 to 0.74, well below SHAP-LightGBM's 0.66 to 1.25. Read together, the two metrics show that LIME's importance magnitudes barely move across bootstrap iterations even as its feature rankings collapse to statistical noise. This is only possible when importance values across features sit tightly clustered together, so that a small perturbation is enough to reorder the ranking without shifting the values by much. CV and SRA capture two distinct dimensions of stability, and relying on either one alone would understate how unstable LIME becomes under class imbalance.

Figure 5

SHAP and LIME Coefficient of Variation (CV) Across Imbalance Levels, Taiwan Dataset



4.4 | Cross-Country Comparison

Explanation stability does vary by country. Table 8 and Figure 6 show the average SHAP and LIME stability compared with different models for Taiwan and the USA datasets. The USA performed significantly better compared to Taiwan. The average SHAP SRA gap ranges from 0.11 to 0.26 for Random Forest and LightGBM, respectively, and at the individual-configuration level it is largest: when the default rate is 1%, the gap is 0.59 for LightGBM, and it tightens as imbalance reduces. A similar pattern is observed by LIME but more severely: at a 1% default rate, the SRA falls to near-zero SRA (0.04 to 0.49) for all three Taiwan models, while for the USA it remains moderate between 0.43 and 0.56 at 1%, to near-identical rankings at 50% imbalance.

Table 8

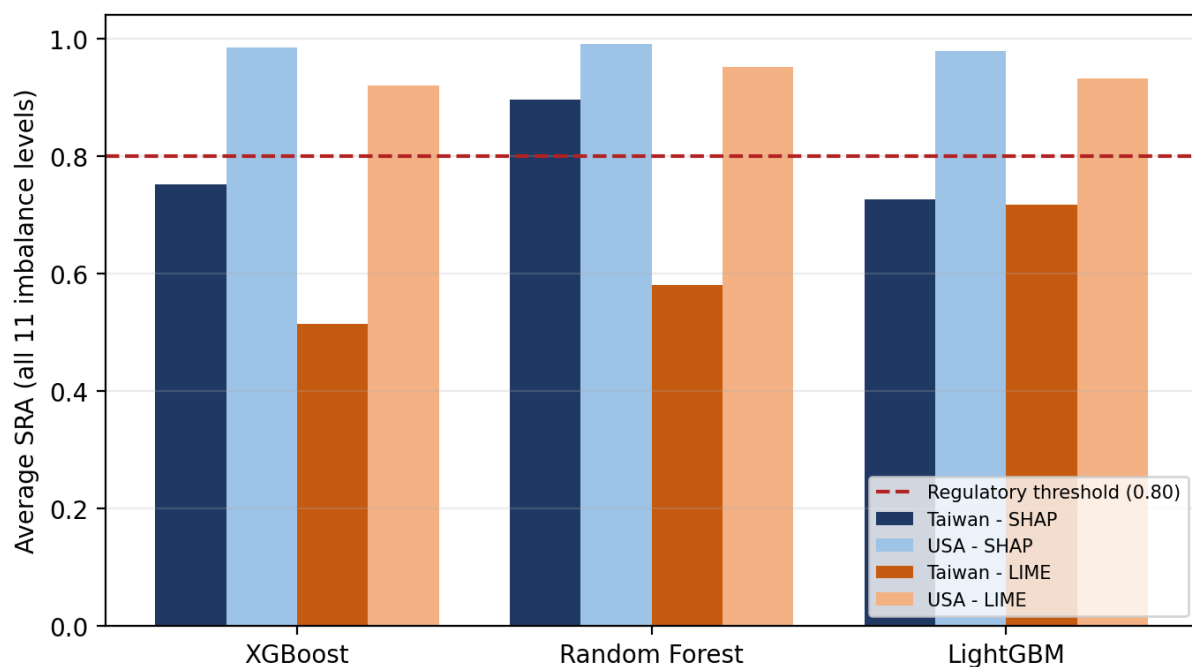
Average SHAP/LIME Stability

Dataset	Model	SHAP SRA	LIME SRA	Above Threshold
Taiwan	XGBoost	0.776	0.490	Neither
Taiwan	Random Forest	0.877	0.744	SHAP
Taiwan	LightGBM	0.714	0.608	Neither
USA	XGBoost	0.978	0.790	SHAP
USA	Random Forest	0.987	0.745	SHAP
USA	LightGBM	0.978	0.780	SHAP

Note. Threshold = 0.80.

Figure 6

Average SHAP and LIME Stability by Model, Taiwan Versus USA



The cross-country difference can be explained by the characteristics of the dataset, not as a property of SHAP or LIME. The benefit of the LendingClub dataset was a well-spread observation period, a diverse borrower population, and a feature set built on standard credit bureau reporting. In combination, all these factors produce a much more stable feature-importance landscape, which across bootstrap iterations reduced the variation.

One reading of this pattern is worth stating carefully, as a hypothesis rather than a finding: post-hoc explanation methods may perform more reliably in mature, data-rich credit markets with standardized bureau reporting than in transitional ones, since SHAP and LIME both depend on having enough well-distributed data to estimate stable local structure. This study cannot test that hypothesis directly. Two datasets from two countries are a comparison of dataset characteristics, not a controlled test of market maturity, and the two also differ in time period, product type, and sample size in ways this design cannot separate from maturity itself. Confirming or rejecting the hypothesis would need multiple datasets from markets at differing stages of development, held constant on the other dimensions, which is a direction for future research rather than a conclusion this study can support.

6 | DISCUSSION

Prior to addressing each research question in detail, throughout this section one framing point matters: the 5% default rate that recurs below is a pattern observed in this study and in [Chen et al. \(2024\)](#), not a universal property of imbalanced classification. The point at which any given model-method combination crosses the 0.80 threshold varies considerably: LightGBM-LIME clears it by 15% on Taiwan, Random Forest-LIME not until 35%, and SHAP on the USA dataset never drops below the threshold at any tested level. Treating 5% as a fixed cutoff would overstate what a single default rate can tell a practitioner across different models, methods, and credit markets; it is a useful rule of thumb for this study's specific conditions, not a general law.

The answer to RQ1 is yes, but the severity differs in both markets. Taiwan dataset is the more extreme case: the huge drop at the 1% default rate to 0.595 for XGBoost in SHAP SRA values and 0.360 for LightGBM. Any post-hoc explanation generated at these levels would be highly unreliable for a credit decision. LIME failed more fundamentally: its rank agreement collapsed to near-zero at 1% for all three models, with the confidence interval for XGBoost and LightGBM straddling zero, meaning the ranking is statistically indistinguishable from having no relationship at all to the balanced-reference ranking.

RQ2 is also supported, though the CV evidence points to a different mechanism than SRA. As Table 7 shows, CV rises rather than falls as the default rate increases, which reflects genuine variation across bootstrap samples in a more balanced training set rather than a loss of stability. The more diagnostic evidence for RQ2 is comparative: LIME's CV remains lower than SHAP's at every imbalance level even as its SRA turns negative, showing that value stability and rank-order stability diverge sharply for LIME under extreme imbalance.

A question closely related to this is whether explanation instability simply tracks weaker predictive performance, instead of being a property of the explanation method itself. The most consistent model for both SHAP and LIME on the Taiwan dataset is Random Forest; however, per-model discrimination scores are not reported in this paper at the level of detail required to test whether the stability reflects Random Forest attaining

higher discrimination, or if it reflects its decision-boundary structure independently of overall performance. This would require matching models on discrimination performance rather than on architecture; specifically, this can be a direction for future work.

The answer to *RQ3* is yes. The direction of the country effects was consistent across both XAI methods: the USA market outperformed Taiwan in SHAP and LIME explanations at every imbalance level. This gap is attributable to the large dataset scale, feature standardization, and the maturity of US credit bureau reporting, not to a property of LIME or SHAP.

First, the threshold of below 5% instability replicated in this study and has held. When analyzing the Taiwan dataset, the SRA for SHAP fell below 0.80 at the 1% level for all three models and remains below the threshold until 5% defaults for XGBoost and LightGBM. When analysing the USA dataset, LIME SRA fell below the threshold at the 1% level. The exact value after which the explanations become unstable may vary by dataset and model; however, the general finding of [Chen et al. \(2024\)](#) that extreme imbalance produces regulatory-grade instability is held true across two different credit markets and three machine learning models.

Other than extending [Chen et al. \(2024\)](#) findings, we documented two new findings. First, at extreme imbalance, LIME's rank agreement breaks down to statistical noise rather than gradual degradation; for two of the three Taiwan models, the confidence intervals straddle zero at 1%. Second, at the 1% default rate, LightGBM's SHAP stability can reach as low as 0.360, which is lower than either model reported by Chen et al. For financial institutions already running LightGBM in production for credit scoring, this finding carries direct relevance.

6.1 | Theoretical Implications

Statistical Learning Theory's core claim is that estimation error grows as the effective sample size within a region of feature space shrinks; the theory itself does not specify a direction for that error. This study extends it empirically by showing the direction differs systematically by explanation mechanism, not just the magnitude; this study also shows that sparsity-driven instability depends on the explanation method. The marginal contribution estimates of SHAP collapse to unreliable but non-directional rankings. LIME's local linear approximation collapses even further under the same sparsity, to the point where its rank agreement with the balanced-reference ranking becomes statistically indistinguishable from zero for two of three Taiwan models at 1% default rate. Empirically, this severity of collapse has not been found in the prior literature. The findings are also in line with Algorithmic Accountability Theory: an audit trail cannot be supported by an explanation that flips direction across runs. They also support Trust Calibration Theory. When a rejected applicant receives a different explanation every time, they would be misled about the actual decision, even though the model's underlying prediction remains the same.

Two of these effects should not be warped into one another. What is measured by this study is computational instability: if the training data changes, would the same model, in a similar case, produce a different feature ranking? Whether instability actually damages a borrower's or regulator's trust is a separate, human question this study does not test; Trust Calibration Theory motivates why the computational finding should matter, but the trust effect itself would need its own study, for example one that shows explanations to real borrowers or compliance reviewers and measures their confidence directly.

Information Asymmetry Theory assumes that lenders use model explanations to minimize the information gaps with borrowers. The findings suggest that once the default rate falls below 5%, SHAP and LIME explanations become noise instead of useful signals. The feature ranking changes significantly when the training data is different despite the same borrower profile, the same model, and different training data composition. An explanation built on such systems reduces the information asymmetry.

6.2 | Practical Implications

The below-mentioned distinctions are practical and regulatory, not purely statistical: as reported in Table 2, a Friedman test at each imbalance level (significant at $p < .05$ everywhere, except the reference level); however, statistical significance and regulatory relevance are two different questions. For example, a drop from 0.87 to 0.75 SRA may be statistically significant without confirming the Friedman test individually at each level; what practically makes it relevant here is that 0.75 is below the regulatory-grade threshold of 0.80 and 0.87 does not; this distinction matters for compliance regardless of significance level.

- For financial institutions: below 5% default rates, LIME is too unstable to generate compliance-grade explanations under GDPR Article 22 or the EU AI Act. SHAP, when it is paired with Random Forest, is the most stable option, especially when default rates are low, on datasets resembling Taiwan.
- For institutions using LightGBM: at low default rates, LightGBM, when paired with SHAP, produces explanations that require closer scrutiny compared to explanations generated from XGBoost or Random Forest. This study has recorded the lowest SHAP stability value with LightGBM.
- For regulators: current GDPR and EU AI Act requirements only demand an explanation; however, they do not require it to be stable. Incorporating metrics such as SRA, or an equivalent stability measure, as part of model validation would close this gap.
- For lower-data-volume credit markets: the large stability gap between Taiwan and the USA shows that explanation stability results do not generalize across credit systems. Institutions should test explanation stability in their own market, rather than assume results obtained in a different market.

6.3 | Limitations and Future Research

Two constraints shape interpretation. The Taiwan dataset covers a single six-month window in 2005, whereas the LendingClub dataset covers 2007-2018. As the two datasets are from different time periods, only the pattern of stability degradation can be compared across countries, not absolute stability values. Both of the datasets also focus on specific product types: credit cards for Taiwan and personal loans for the USA. The stability pattern may differ for mortgages and credit cards. Second, this study is restricted to tree-based ensembles. Results may not remain the same for neural networks or linear models; this remains an open question for future studies. Future work should extend this design to crisis-period data to stress-test stability, additional national credit markets, and interpretable models as a comparison point against post-hoc explanation, and could reintroduce macroeconomic covariates once a genuine period of economic variation is available to test against.

The confidence intervals reported throughout this paper (Tables 3 through 6) come from $B = 50$ bootstrap iterations per imbalance level, a fully seeded and reproducible protocol (seed = 42), computed on a subsample of 50 target observations rather than the originally planned 200. This is fewer iterations and a smaller target sample than the $B = 100$ iterations and 200-observation target set described in the original design (Section 4.3). The wider confidence intervals visible in Tables 3 through 6 reflect this reduced scale directly; a full $B = 100$, 200-observation run would be expected to narrow these intervals without necessarily changing the reported means, but this has not been tested and is a direction for future work.

A further limitation concerns model configuration. Hyperparameters for all three models were fixed rather than tuned via a formal search, and two different fixed configurations were used across the analysis: one for the discrimination scores in Section 5.1, and each library's default configuration for the models that generated the SHAP and LIME explanations reported throughout the remainder of the paper. Model discrimination is also evaluated inconsistently across countries: a single held-out split for Taiwan, and 5-fold cross-validated AUC for the USA dataset. Neither inconsistency affects the explanation-stability findings that are the core contribution of this study, since those results come from the same models throughout each dataset, but both should be resolved through a unified, tuned pipeline in future work.

A further boundary concerns how imbalance was constructed: every training set used under-sampling of the majority class only, with no comparison against class-weighting or synthetic oversampling. It is possible that the instability patterns reported here are specific to undersampling rather than a general property of class imbalance; this has not been tested and is named as a direction for future work below.

7 | CONCLUSION

This paper replicates and extends [Chen et al. \(2024\)](#) using open data. It extends their work by adding a third machine learning model and a second credit market. The main finding holds in both settings: as class imbalance increases, SHAP and LIME explanation stability decreases, with the sharpest degradation below a 5% default rate. Two results are new to the literature. The first is that LIME's rank agreement collapses to statistical noise at extreme imbalance on the Taiwan dataset, rather than degrading gradually. The second is the consistent stability advantage of the more data-rich, standardized USA market, across both explanation methods. Explanations generated at low default rates should be treated as provisional by financial institutions and regulators, especially those relying on SHAP or LIME to meet transparency requirements under GDPR or the EU AI Act. Stability testing should be adopted as a standard component of model validation.

The results demonstrate, for the Taiwan dataset specifically, that SHAP and LIME feature-ranking stability falls below the 0.80 regulatory-grade threshold at low default rates for all three models tested, with SHAP SRA as low as 0.360 (LightGBM, 1%) and LIME SRA collapsing to near-zero (0.038 to 0.487) across all three models at 1% (Tables 3 and 5). They demonstrate that this pattern is weaker but still present on the USA LendingClub dataset, where LIME does not clear the threshold until 30-35% (Table 6), while SHAP remains above the threshold throughout (Table 4). They demonstrate that CV moves in the opposite direction to SRA as balance improves, which reflects genuine bootstrap variation rather than instability (Table 7). What the results do not demonstrate, because the design does not test it, is whether these patterns hold for any model

beyond the three tree-based ensembles used here, or whether they would replicate under a formally tuned hyperparameter search.

This study had three claims of originality, consistent with the abstract: it is the first open-data, cross-country replication of [Chen et al. \(2024\)](#); it is the first study to report the collapse of LIME's feature-ranking agreement to statistical noise under extreme class imbalance, rather than gradual degradation; and it is the first to test explanation stability of LightGBM together with XGBoost and Random Forest in this setting, where it produced the lowest SHAP stability value recorded in the study. A fourth, secondary contribution is related to methodological rather than empirical: a disparity between the models used for discrimination scoring and those used for explanation generation; the paper concealed and reported an honest performance metric (the Gini coefficient) instead of an unimplemented cost-weighted H-measure.

Based on the limitations discussed above, three extensions follow. First, under-sampling should be tested against class-weighting and synthetic oversampling on at least a subset of configurations, to establish whether the instability patterns reported here are specific to how imbalance was operationalized. Second, the cross-country design should be extended to additional credit markets and to crisis-period data, to test whether the 5% pattern observed here holds under macroeconomic conditions this study's stable six-month window could not capture. Third, a single, properly tuned hyperparameter configuration should be used consistently across both the discrimination-scoring and explanation-generating stages, alongside a correctly implemented cost-weighted H-measure; this study used two different configurations for these two purposes, disclosed as a limitation in Section 9, and unifying them is a natural next step. The confidence intervals and Friedman/Wilcoxon significance tests called for in earlier drafts of this future-research list have already been carried out and are reported in Table 2 and throughout Section 5.

Acknowledgment: The authors would like to express their sincere thanks to the editor and the anonymous reviewers for their helpful comments and suggestions.

Author Contributions: Conceptualization: Tufail, A. and Sohail, M.K.; Data curation: Tufail, A.; Formal analysis: Tufail, A. and Sohail, M.K.; Investigation: Tufail, A.; Methodology: Tufail, A. and Sohail, M.K.; Supervision: Sohail, M.K.; Validation: Tufail, A. and Sohail, M.K.; Visualization: Tufail, A.; Writing – original draft: Tufail, A.; Writing – review & editing: Tufail, A. and Sohail, M.K. Both authors approved the final version of the manuscript for publication.

Declaration of Conflicting Interest: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Licenses and Copyright: This is an open-access article, free of all copyright, and fulfills the DOAJ definition of open access. This work is licensed under a “[Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)”. This permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Data Availability Statement: The data that support the findings of this study are openly available. The Taiwan Credit Card Dataset is available from the UCI Machine Learning Repository at <https://archive.ics.uci.edu/dataset/350>. The LendingClub USA Dataset is available via Zenodo at <https://zenodo.org/records/11295916>.

Plagiarism Statement: This article was screened using Turnitin. No plagiarism was detected. The Overall Similarity Index (SI) and Single Source Similarity are within the Higher Education Commission (HEC) prescribed thresholds (Overall SI <19%; Single Source <5%).

Artificial Intelligence (AI) Statement: This article was screened using Turnitin's AI writing detection tool to identify AI-generated and AI-paraphrased content. The AI-generated content is within the journal's prescribed threshold and complies with the HEC Anti-Plagiarism and AI Policy.

Disclaimer/Publisher’s Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher “AGASR” and/or the editor(s). The Publisher AGASR and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

REFERENCE

- Ariza-Garzón, M. J., Sanz-Guerrero, M., & Arroyo Gallardo, J. (2024). Lending club loan dataset for granting models. Lending Club loan dataset for granting models. <https://doi.org/https://doi.org/10.5281/zenodo.11295916>
- Ballegeer, M., Bogaert, M., & Benoit, D. F. (2025). Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring. *European Journal of Operational Research*, 326(3), 630-640. <https://doi.org/10.1016/j.ejor.2025.05.039>
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32. <https://doi.org/https://doi.org/10.1023/A:1010933404324>
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with applications*, 39(3), 3446-3453. <https://doi.org/10.1016/j.eswa.2011.09.033>
- Bueff, A. C., Cytryński, M., Calabrese, R., Jones, M., Roberts, J., Moore, J., & Brown, I. (2022). Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. *Expert systems with applications*, 202, 117271. <https://doi.org/10.1016/j.eswa.2022.117271>
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management: N. Bussmann et al. *Computational Economics*, 57(1), 203-216. <https://doi.org/https://doi.org/10.1007/s10614-020-10042-0>
- Chan, C.-P., Tsai, C.-H., Tang, F.-K., & Yang, J.-H. (2025). A SHAP-based comparative analysis of machine learning model interpretability in financial classification tasks. *Journal of Applied Economic Sciences*, 20(3), 385-400. [https://doi.org/https://doi.org/10.57017/jaes.v20.3\(89\).03](https://doi.org/https://doi.org/10.57017/jaes.v20.3(89).03)
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357. <https://doi.org/https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,
- Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, 312(1), 357-372. <https://doi.org/10.1016/j.ejor.2023.06.036>
- Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, 106263. <https://doi.org/10.1016/j.asoc.2020.106263>
- Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62. <https://doi.org/10.1145/2844110>
- Dong, B., Zhang, X., Hong, Z., Yan, C., Zhu, W., Hou, L., & Feng, Y. (2026). Explainable Deep Tabular Learning for Credit Risk Assessment: An Information-Theoretic Cross-Attentional Transformer Approach. *Entropy*, 28(8), 837. <https://doi.org/https://doi.org/10.3390/e28080837>
- Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International conference on intelligent computing,
- Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1), 103-123. <https://doi.org/https://doi.org/10.1007/s10994-009-5119-5>
- He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence),
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Kozodoi, N., Lessmann, S., Alamgir, M., Moreira-Matias, L., & Papakonstantinou, K. (2025). Fighting sampling bias: A framework for training and evaluating credit scoring models. *European Journal of Operational Research*, 324(2), 616-628. <https://doi.org/10.1016/j.ejor.2025.01.040>
- Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in artificial intelligence*, 5(4), 221-232. <https://doi.org/https://doi.org/10.1007/s13748-016-0094-0>
- Lazar, A., Pokhrel, P., & Das, S. (2026). Beyond accuracy: A comprehensive comparative study of gradient boosting versus tabular deep learning and explainability techniques for mixed-type tabular data models using SHAP and LIME. *International Journal on Artificial Intelligence*

- Tools, 35(02), 2640003. <https://doi.org/https://doi.org/10.1142/S0218213026400038>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80. <https://doi.org/https://doi.org/10.1518/hfes.46.1.50.30392>
- LendingClub Dataset. (2007-2018). *LendingClub Dataset Full 2007 to 2018* (Kaggle). <https://www.kaggle.com/datasets/panchamhah/to/lendingclub-dataset-full-2007-to-2018>
- Liu, W., Fan, H., & Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert systems with applications*, 189, 116034. <https://doi.org/10.1016/j.eswa.2021.116034>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions Advances in neural information processing systems,
- Machado, M. R., & Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert systems with applications*, 200, 116889. <https://doi.org/10.1016/j.eswa.2022.116889>
- Misheva, B. H., Osterrieder, J., Hirs, A., Kulkarni, O., & Lin, S. F. (2021). Explainable AI in credit risk management. *arXiv preprint arXiv:2103.00949*. <https://doi.org/https://doi.org/10.48550/arXiv.2103.00949>
- Moscato, V., Picariello, A., & Sperli, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert systems with applications*, 165, 113986. <https://doi.org/10.1016/j.eswa.2020.113986>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, California, USA.
- Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American economic review*, 71(3), 393-410. <https://doi.org/https://www.jstor.org/stable/1802787>
- Vapnik, V. N. (1998). *Statistical learning theory {Adaptive and learning systems for signal processing, communications, and control}*. Wiley and Sons.
- Visani, G., Bagli, E., Chesani, F., Poluzzi, A., & Capuzzo, D. (2022). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1), 91-101. <https://doi.org/https://doi.org/10.1080/01605682.2020.1865846>
- Wang, Z., Tian, Y., Li, S., & Xiao, J. (2025). A secure cross-silo collaborative method for imbalanced credit scoring. *European Journal of Operational Research*, 326(2), 357-373. <https://doi.org/10.1016/j.ejor.2025.04.020>
- Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2), 2473-2480. <https://doi.org/10.1016/j.eswa.2007.12.020>
- Zhu, J., Zhang, M., Zhang, X., & Yu, L. (2026). Triplet guided dual-discriminator filtering network for extremely imbalanced credit classification. *International Review of Financial Analysis*, 105474. <https://doi.org/https://doi.org/10.1016/j.irfa.2026.105474>