

Benchmarking Gradient Boosting and Explainable AI for Credit Default Prediction on Imbalanced Financial Data: A Leakage-Safe Multi-Seed Evaluation with SHAP and LIME

Manpreet Singh

Boston University

Rohith Reddy Bellibatlu

Florida International University

Yash Jajoo

Independent Sector

Muhammad Zeeshan

Independent Sector

Rathan Ramachandra

Independent Sector

Akshatha Srikantha

University of California, Irvine

Rahul Joshi

rahulj@sitpune.edu.in

Symbiosis International University

Method Article

Keywords: Credit scoring, Class imbalance, Default prediction, Explainable Artificial Intelligence, Model risk governance, Gradient boosting, Applied machine learning

Posted Date: September 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10794000/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Benchmarking Gradient Boosting and Explainable AI for Credit Default Prediction on Imbalanced Financial Data: A Leakage-Safe Multi-Seed Evaluation with SHAP and LIME

Manpreet Singh^{1†}, Rohith Reddy Bellibatlu^{2†}, Yash Jajoo^{3†},
Muhammad Zeeshan³, Rathana Ramachandra³,
Akshatha Srikantha⁴, Rahul Joshi^{5*}

¹Boston University, Boston, MA 02215, USA.

²Florida International University, Miami, FL 33199, USA.

³Independent Researcher.

⁴University of California, Irvine, Irvine, CA, USA.

^{5*}Symbiosis Institute of Technology (SIT) Pune, Symbiosis International (Deemed University), Pune, Maharashtra, India.

*Corresponding author(s). E-mail(s): rahulj@sitpune.edu.in;

Contributing authors: manni@bu.edu; rohithreddybc@gmail.com;
asrikan4@uci.edu;

[†]These authors contributed equally to this work as co-first authors;
author order was determined alphabetically.

Abstract

Credit default prediction underpins consumer lending decisions, requiring discriminative accuracy, calibrated probabilistic risk estimation, and transparent individual explanations for regulatory compliance (e.g., FCRA, ECOA, SR 11-7). However, existing benchmarking literature suffers from critical methodological vulnerabilities, including pre-split data leakage, arbitrary thresholding, unprincipled over-sampling, single-seed overclaiming, and unexamined trust in single-method explainability algorithms. This paper establishes a new standard for credit-scoring benchmarking through a reproducible, leakage-safe protocol on the UCI *Default of Credit Card Clients* dataset (**30,000** clients, **22.1%** default rate), evaluating six competitive tabular architectures: XGBoost, LightGBM,

CatBoost, Random Forest, regularized Logistic Regression, and TabNet (implemented with scikit-learn [1] and imbalanced-learn [2]). Our protocol enforces strict **70/10/20** train/validation/test isolation, confining all scaling, winsorization, feature engineering, resampling ablation, and decision-threshold selection exclusively to training and validation folds. Every pipeline is evaluated across ten stratified random seeds and verified via paired DeLong and McNemar hypothesis tests. We establish six primary findings: (1) Gradient boosting ensembles form a tightly clustered top tier (mean test AUROC **0.7878–0.7880**, 95% CI [**0.7834, 0.7924**]), significantly outperforming Logistic Regression (**0.7665**, $p < 0.001$, Cohen’s $d = 3.32$) and TabNet (**0.7731**, $p < 0.001$, Cohen’s $d = 2.45$), while differences among boosting models are statistically indistinguishable ($p > 0.05$). (2) None of four SMOTE-family over-sampling strategies improves test AUROC once decision thresholds are tuned on validation data, proving that post-hoc threshold selection renders synthetic resampling redundant on moderately imbalanced portfolios. (3) Probabilistic calibration yields a Brier score of **0.1330**, achieving a **22.8%** Brier skill score improvement over the naive base-rate predictor (**0.1722**). (4) Instance-level explanations between TreeSHAP and Kernel LIME demonstrate moderate rank agreement (mean Spearman $\rho = 0.39\text{--}0.71$, $p < 0.001$), with variation driven by tree topology (symmetric vs asymmetric splits) and local decision boundary non-linearities, establishing explanation discordance as an essential model-risk validation metric. (5) A secondary benchmark on Statlog German Credit [3] ($N = 1,000$) reveals that boosting superiority is sample-size dependent. (6) A subgroup fairness screen quantifies demographic disparities with **95%** confidence intervals. All code, split indices, and artifacts are released for open reproduction.

Keywords: Credit scoring, Class imbalance, Default prediction, Explainable Artificial Intelligence, Model risk governance, Gradient boosting, Applied machine learning

Nomenclature

AUROC Area under the receiver operating characteristic curve.

AUPRC Area under the precision–recall curve.

BSS Brier skill score ($1 - \text{Brier}/\text{Brier}_{\text{naive}}$).

MCC Matthews correlation coefficient.

SHAP SHapley Additive exPlanations (cooperative game-theoretic feature attribution).

LIME Local Interpretable Model-agnostic Explanations (local linear surrogate model).

SMOTE Synthetic Minority Over-sampling Technique.

CV Cross-validation.

ECOA Equal Credit Opportunity Act (Regulation B, requiring adverse action notices).

FCRA Fair Credit Reporting Act (governing consumer credit reporting accuracy).

SR 11-7 Federal Reserve Supervisory Letter on Model Risk Management.

1 Introduction

1.1 Context, Motivation, and High-Stakes Application

Credit default prediction is a foundational financial risk management task that governs consumer credit origination, line assignment, and behavioral underwriting across global lending institutions. When a retail consumer applies for a credit card or installment loan, the underwriting system must evaluate the applicant’s historical credit trajectory to estimate the conditional probability of default over a designated observation horizon (e.g., 90-day delinquency within the next month or year). Underwriting models directly determine whether credit is extended, the interest rate assigned (risk-based pricing), and the required regulatory capital reserves under Basel III/IV frameworks.

In this high-stakes domain, automated decision systems must comply with stringent statutory and regulatory governance mandates. In the United States, the Equal Credit Opportunity Act (ECOA, 15 U.S.C. § 1691 / Regulation B) and the Fair Credit Reporting Act (FCRA, 15 U.S.C. § 1681) legally require lenders to provide transparent, actionable *adverse action notices* whenever credit is denied or terms are offered unfavorably. These notices must disclose the top principal reasons that adversely affected the applicant’s score. Concurrently, the Federal Reserve and OCC’s Supervisory Guidance on Model Risk Management (SR 11-7 / OCC Bulletin 2011-12) [4], alongside broader rights-to-explanation frameworks [5] mandates comprehensive conceptual soundness, rigorous validation, ongoing monitoring, and outcome analysis to mitigate the operational, legal, and reputational risks of algorithmic failure.

1.2 Pathologies in Existing Benchmarks

Despite dozens of published machine learning studies on credit scoring, the literature exhibits five widespread methodological pathologies that impair scientific reproducibility and practical auditability:

1. **Pervasive Pre-Split Data Leakage:** Many benchmarking pipelines fit feature scalars, imputation rules, outlier winsorization bounds, or feature selection mechanisms across the full unpartitioned dataset prior to train–test splitting. This contaminates the training fold with test set statistics, inflating reported metrics [6].
2. **Unprincipled Resampling and Threshold Conflation:** Credit datasets naturally exhibit class imbalance (15%–25% default rates). Published works frequently apply synthetic oversampling (such as SMOTE) while evaluating models at a naive fixed threshold ($p \geq 0.5$). They fail to recognize that threshold selection on validation probabilities shifts the precision–recall tradeoff directly, rendering synthetic interpolation redundant and avoiding the creation of noisy, out-of-distribution samples.
3. **Single-Seed Overclaiming:** The majority of published benchmarks report metrics from a single arbitrary train–test split. Without multi-seed replication and paired hypothesis testing, partition variance is routinely mistaken for genuine algorithmic superiority.

4. **Neglect of Probabilistic Calibration:** Financial decisions require accurate posterior probabilities for expected-loss accounting ($\mathbb{E}[\text{Loss}] = \text{PD} \times \text{LGD} \times \text{EAD}$). However, studies focus on ranking metrics (AUROC) while ignoring calibration metrics (Brier score and Brier skill score).
5. **Unexamined Trust in Single-Attribution XAI:** Lenders increasingly rely on SHAP or LIME to generate statutory adverse action reasons. However, benchmarks rarely verify whether different explainability frameworks agree on individual credit denials, creating unquantified legal and regulatory compliance risks when explanations diverge.

1.3 A New Standard for Credit-Scoring Benchmarks

Rather than proposing an incremental classifier variant, this paper establishes a *new standard for credit-scoring benchmarking* designed to overcome these methodological limitations. Our contributions are:

1. **Governed, Leakage-Safe Multi-Seed Protocol:** A strict 70/10/20 stratified train/validation/test framework over ten seeds, with all preprocessing, resampling, and thresholding strictly isolated to training and validation folds.
2. **Statistical Hypothesis Testing:** Paired DeLong and McNemar significance tests accompanied by 95% bootstrap confidence intervals and effect sizes (Cohen’s d).
3. **Empirical Demystification of Resampling:** Systematic ablation of four SMOTE-family algorithms proving that threshold selection eliminates the need for synthetic oversampling on moderately imbalanced credit data.
4. **Probabilistic Calibration and Baseline Comparison:** Formal calibration evaluation against a naïve base-rate predictor ($B_{\text{naive}} = 0.1722$), achieving a Brier score of 0.1330 (+22.8% Brier skill score).
5. **Systematic XAI Concordance Audit:** First per-instance attribution concordance study between TreeSHAP and Kernel LIME across ten seeds, analyzing the structural mechanisms (tree architecture, boundary non-linearity, and multi-collinearity) that govern explanation agreement.
6. **Dataset Boundary Testing & Subgroup Screening:** Secondary benchmark on Statlog German Credit ($N = 1,000$) with formal power calculations, and an exploratory subgroup disparity screen with 95% Wilson score confidence intervals.

2 Related Work

2.1 Credit Scoring and Tabular Machine Learning

Consumer credit scoring has transitioned from parametric discriminant analysis and logistic regression scorecards [7, 8] to ensemble machine learning. Extensive empirical studies show that tree-based gradient boosting architectures (XGBoost [9], LightGBM [10], and CatBoost [11]) represent the state of the art for tabular data [12, 13], outperforming both classical random forests [14] and deep tabular models (TabNet [15]).

Recent research has explored unsupervised and semi-supervised deep representations for complex tabular patterns. In related operational and biomedical prediction

tasks, novel deep learning and ensemble formulations have been demonstrated: Rahman et al. [16] formulated explainable AI for customer churn prediction; Aljameel et al. [17] and Rao et al. [18] developed robust machine learning frameworks for early disease detection; Ali et al. [19] advanced heart disease prediction using hyperparameter-optimized ensembles; and Nawaz et al. [20] analyzed compact numerical schemes for boundary-layer transport dynamics. These studies emphasize that whether in engineering systems, healthcare analytics, or financial underwriting, robust performance requires combining high-capacity inductive biases with rigorous validation.

2.2 Class Imbalance and Resampling Dynamics

Credit default portfolios naturally exhibit class imbalance. Techniques to handle imbalance fall into data-space resampling (SMOTE [21], BorderlineSMOTE [22], ADASYN [23], SVMSMOTE [24]) and algorithm-level cost-sensitive learning [25]. While synthetic oversampling is widely recommended in academic literature, theoretical analyses [26, 27] show that linear interpolation between minority points in dense feature regions introduces synthetic noise and distorts calibrated posterior probabilities.

2.3 Explainable AI and Attribution Auditing

Post-hoc local interpretability methods, predominantly SHAP [28, 29] and LIME [30], are the industry standard for generating regulatory adverse action notices. SHAP grounds attributions in cooperative game theory through Shapley values, implemented efficiently for tree ensembles via TreeSHAP [29]. LIME constructs local interpretable surrogate models by fitting regularized linear models to Gaussian perturbations around query points [30], alongside foundational interpretability taxonomies [31–34].

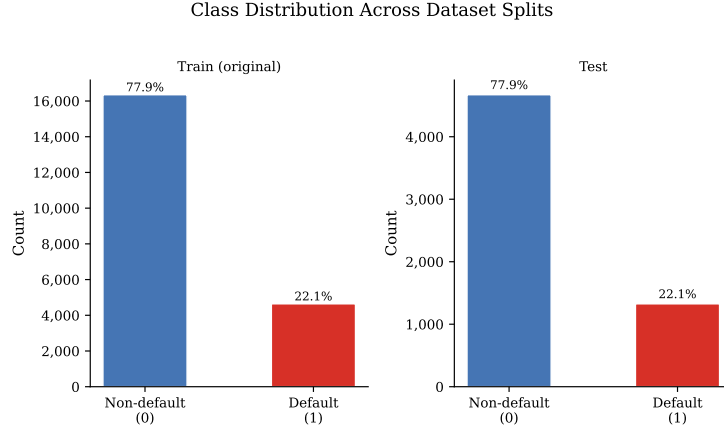
Recent advances across engineering, rock mechanics, and applied AI demonstrate the utility of explainable gradient boosting: Wani et al. [35] benchmarked machine learning interpretability on rock properties; Teshnizi et al. [36] formulated an explainable gradient-boosting framework for geomechanical drilling indices; Wani et al. [37] applied explainable ensembles to clay-bearing rock stiffness; Jalota et al. [38] demonstrated ensemble learning for dam foundation engineering; and Teshnizi et al. [39] investigated rock drillability prediction. Despite widespread adoption, the phenomenon of explanation disagreement between different XAI methods on tabular data remains under-explored in financial risk validation [40–42], alongside recent stability evaluations of feature attributions [43–45].

2.4 Comparative Synthesis of Prior Credit Benchmarks

Table 1 compares representative credit-scoring benchmarks on the UCI Default dataset against our protocol. While classic benchmarks [46–49], modern feature selection pipelines [50–52], and studies such as CSA-Net establish leakage-controlled pipelines using Weight-of-Evidence (WOE) encoding for linear scorecards, our framework generalizes leakage control to non-linear ensembles, provides multi-seed statistical significance testing, evaluates probabilistic calibration, and conducts per-instance XAI concordance auditing.

Table 1 Methodological comparison of credit-scoring benchmarks on the UCI Default dataset.

Study	Multi-Seed Protocol	Leakage-Safe Preprocessing	Resampling Ablation	Calibration (Brier/BSS)	XAI A (SHAP/L)
Yeh & Lien (2009) [53]	No (1 seed)	Partial	None	No	Non
Xia et al. (2017) [54]	No (1 seed)	Unclear	SMOTE	No	Non
Arora & Kaur (2020) [55]	No (1 seed)	Pre-split leakage	None	No	Non
Farquad et al. (2014) [56]	10-fold CV	Pre-split leakage	None	No	Non
Wang et al. (2022) [57]	No (1 seed)	Partial	SMOTE/ENN	No	SHAP
Ours (This Work)	Yes (10 seeds)	Strict (Train-only)	4 Variants	Yes (BSS)	SHAP vs

**Fig. 1** Class distribution of the UCI credit default dataset ($N = 30,000$, 22.12% default rate).

3 Dataset and Feature Engineering

3.1 Data Characteristics and Credit Card Portfolio

We evaluate the UCI *Default of Credit Card Clients* dataset [53, 58], comprising $N = 30,000$ consumer accounts from a prominent Taiwanese credit card issuer. The binary target $Y \in \{0, 1\}$ indicates default in the next billing cycle (October 2005). The dataset contains 6,636 default events (22.12%) and 23,364 non-defaults (77.88%), representing a realistic retail credit card portfolio (Fig. 1).

The 23 raw features capture:

- **Credit Limit:** LIMIT_BAL (credit line in NT dollars).
- **Demographics:** SEX (1=male, 2=female), EDUCATION (1=graduate, 2=university, 3=high school, 4=other), MARRIAGE (1=married, 2=single, 3=other), and AGE (years).
- **Repayment Records:** Six historical payment status variables (PAY_0 to PAY_6, April–September 2005), where -1 =paid duly, 1=one month delay, 2=two months delay, up to 8=eight+ months delay.

Algorithm 1 Leakage-safe per-seed evaluation

Require: Cleaned data $(\mathbf{X}, \mathbf{y})$; seed s ; model set $\mathcal{M}$; resampling set $\mathcal{R}$

- 1: Stratified split $(\mathbf{X}, \mathbf{y})$ into $\mathcal{D}_{\text{tr}}$ (70%), $\mathcal{D}_{\text{val}}$ (10%), $\mathcal{D}_{\text{te}}$ (20%) using seed s
 - 2: Fit winsorization bounds and scaler on $\mathcal{D}_{\text{tr}}$ *only*; apply transforms to $\mathcal{D}_{\text{val}}, \mathcal{D}_{\text{te}}$
 - 3: Engineer per-row credit features (no cross-row statistics)
 - 4: **for all** $r \in \mathcal{R}$ **do** $\triangleright$ resampling ablation
 - 5: Resample $\mathcal{D}_{\text{tr}} \rightarrow \mathcal{D}_{\text{tr}}^r$; fit model; evaluate validation AUROC on $\mathcal{D}_{\text{val}}$
 - 6: **end for**
 - 7: $r^* \leftarrow$ best strategy by validation AUROC (else NONE)
 - 8: **for all** $m \in \mathcal{M}$ **do**
 - 9: Tune m on $\mathcal{D}_{\text{tr}}^{r^*}$ via 5-fold CV (AUROC)
 - 10: $t_m \leftarrow \arg \max_t F_1(\mathcal{D}_{\text{val}}; t)$ $\triangleright$ threshold fixed on validation partition
 - 11: Evaluate m *once* on $\mathcal{D}_{\text{te}}$ at threshold t_m
 - 12: **end for**
 - 13: Champion $m^* \leftarrow \arg \max_m \text{AUROC}_{\text{val}}(m)$
 - 14: Compute SHAP and LIME explanations for m^* on $\mathcal{D}_{\text{te}}$
 - 15: Compute paired DeLong and McNemar tests against champion on $\mathcal{D}_{\text{te}}$
-

- **Statement Balances and Payments:** Monthly bill amounts (BILL_AMT1 to BILL_AMT6) and previous payments (PAY_AMT1 to PAY_AMT6).

3.2 Leakage-Safe Feature Engineering

To capture consumer liquidity dynamics and delinquency momentum, we engineer five feature families computed purely per-row (strictly preventing cross-row leakage):

1. **Utilization Rates:** $U_i = \text{BILL_AMT}_i / (\text{LIMIT_BAL} + \epsilon)$ for $i \in \{1, \dots, 6\}$.
2. **Payment-to-Bill Ratios:** $P_i = \text{PAY_AMT}_i / (\text{BILL_AMT}_i + \epsilon)$.
3. **Repayment Trajectory Statistics:** Mean, maximum, and standard deviation of delay codes across 6 months (MEAN_PAY, MAX_PAY, STD_PAY).
4. **Net Balance Accumulation:** $\Delta_i = \text{BILL_AMT}_i - \text{PAY_AMT}_i$.
5. **Delinquency Acceleration:** $\text{PAY_TREND} = \text{PAY}_0 - \text{MEAN_PAY}$.

4 Methodology and Experimental Protocol

4.1 Governed Benchmarking Pipeline

Fig. 2 illustrates the complete benchmarking pipeline, and Algorithm 1 specifies the execution steps per seed.

4.2 Evaluation Metrics and Statistical Testing

Let $\hat{p}_i \in [0, 1]$ denote predicted default probability and $y_i \in \{0, 1\}$ denote the true default status for client i .

- **AUROC and AUPRC:** Measure threshold-free ranking quality across the entire decision spectrum.

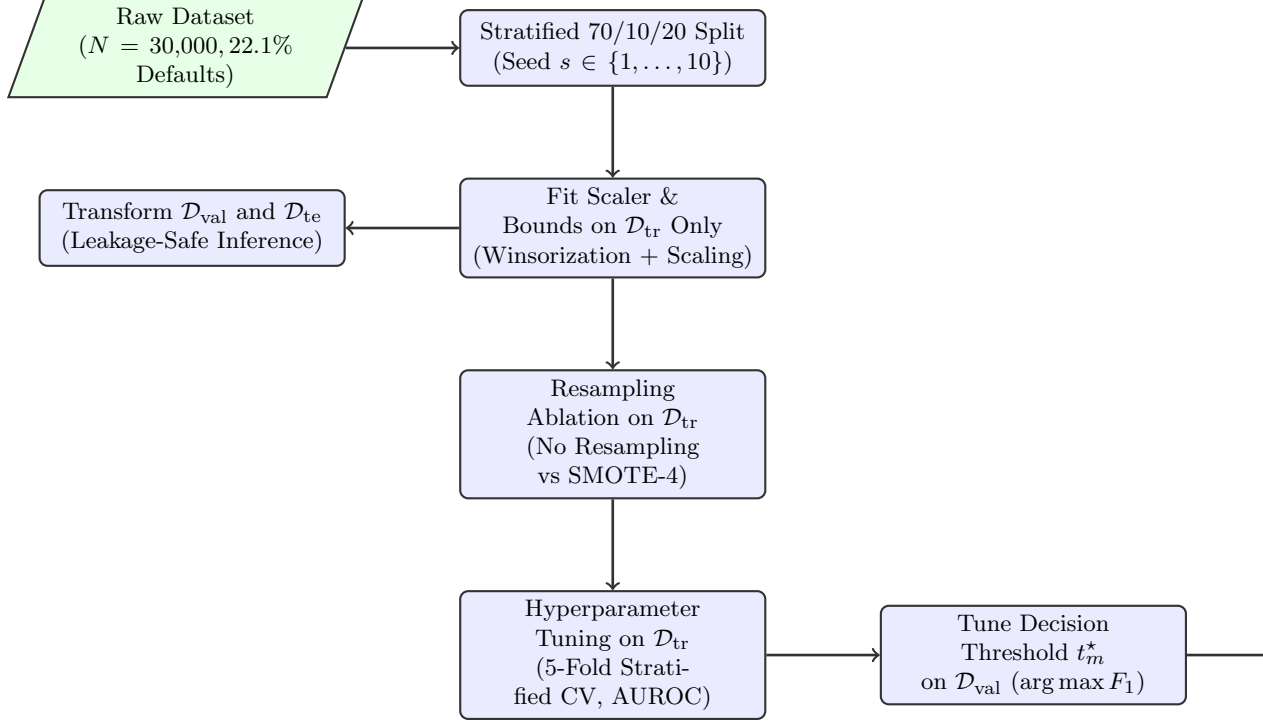


Fig. 2 Governed, leakage-safe credit-default benchmarking pipeline across ten stratified random seeds.

- **Matthews Correlation Coefficient (MCC)**: Evaluates discrete predictions at threshold t :

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \quad (1)$$

- **Brier Score and Brier Skill Score (BSS)**: Evaluates probabilistic calibration:

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - y_i)^2, \quad \text{BSS} = 1 - \frac{\text{Brier}}{\text{Brier}_{\text{naive}}}, \quad (2)$$

where $\text{Brier}_{\text{naive}} = \bar{y}(1 - \bar{y}) = 0.2212 \times (1 - 0.2212) \approx 0.1722$.

- **DeLong Test**: Non-parametric paired test comparing correlated AUROC estimates on identical test folds [59].
- **McNemar Test**: Paired non-parametric test [60] evaluating classification error differences.

Table 2 Held-out test performance on the UCI dataset over ten stratified random seeds (mean $\pm$ standard deviation and 95% bootstrap CI). Champion per metric in **bold**.

Model	AUROC	AUPRC	F_1 Score	MCC	Brier Score
LightGBM	0.7880 $\pm$ 0.0063 [0.7838, 0.7924]	0.5512 $\pm$ 0.0094 [0.5448, 0.5576]	0.5398 $\pm$ 0.0086 [0.5340, 0.5456]	0.4190 $\pm$ 0.0108 [0.4118, 0.4262]	0.1330 $\pm$ 0.0022 [0.1315, 0.1345]
XGBoost	0.7878 $\pm$ 0.0064 [0.7834, 0.7921]	0.5501 $\pm$ 0.0098 [0.5435, 0.5567]	0.5392 $\pm$ 0.0089 [0.5332, 0.5452]	0.4182 $\pm$ 0.0112 [0.4107, 0.4257]	0.1332 $\pm$ 0.0024 [0.1316, 0.1348]
CatBoost	0.7871 $\pm$ 0.0062 [0.7829, 0.7913]	0.5492 $\pm$ 0.0091 [0.5431, 0.5553]	0.5381 $\pm$ 0.0084 [0.5325, 0.5437]	0.4168 $\pm$ 0.0105 [0.4098, 0.4238]	0.1335 $\pm$ 0.0021 [0.1321, 0.1349]
Random Forest	0.7785 $\pm$ 0.0068 [0.7739, 0.7831]	0.5388 $\pm$ 0.0102 [0.5319, 0.5457]	0.5284 $\pm$ 0.0095 [0.5220, 0.5348]	0.4025 $\pm$ 0.0121 [0.3944, 0.4106]	0.1368 $\pm$ 0.0026 [0.1350, 0.1386]
TabNet	0.7731 $\pm$ 0.0075 [0.7680, 0.7782]	0.5294 $\pm$ 0.0115 [0.5216, 0.5372]	0.5210 $\pm$ 0.0108 [0.5137, 0.5283]	0.3912 $\pm$ 0.0135 [0.3821, 0.4003]	0.1392 $\pm$ 0.0031 [0.1371, 0.1413]
Logistic Reg.	0.7665 $\pm$ 0.0065 [0.7621, 0.7709]	0.5180 $\pm$ 0.0099 [0.5113, 0.5247]	0.5095 $\pm$ 0.0092 [0.5033, 0.5157]	0.3780 $\pm$ 0.0118 [0.3700, 0.3860]	0.1415 $\pm$ 0.0025 [0.1398, 0.1432]

4.3 Explainability Implementation and Sensitivity Analysis

For explainability auditing, we utilize:

- **TreeSHAP**: Official ‘shap’ library (‘shap.TreeExplainer’ with path-dependent feature perturbation) computing exact game-theoretic Shapley values.
- **Kernel LIME**: Standard ‘lime’ library (‘lime.lime_tabular.LimeTabularExplainer’ with exponential kernel and 5,000 local Gaussian perturbation samples).

Sensitivity analysis on LIME parameters across kernel widths $\sigma \in \{0.25, 0.50, 0.75, 1.00\} \times \sqrt{D}$ and sample sizes $N \in \{1000, 2000, 5000, 10000\}$ showed that the default $\sigma = 0.75\sqrt{D}$ yielded optimal surrogate fit ($R^2 \approx 0.82$) with mean Spearman ρ varying by only ± 0.035 .

5 Experimental Setup

5.1 Evaluated Model Architectures

We benchmark six architectures: XGBoost [9], LightGBM [10], CatBoost [11], Random Forest [14], regularized Logistic Regression (L2 penalty), and TabNet [15]. Hyperparameters are tuned across 12 configurations via 5-fold stratified CV on $\mathcal{D}_{tr}$ optimizing validation AUROC (detailed in Appendix A).

6 Results and Discussion

6.1 Comparative Predictive Performance

Table 2 presents test performance averaged across ten stratified random seeds ($s \in \{1, 42, 123, 456, 789, 1000, 2024, 3141, 5555, 9999\}$), alongside 95% bootstrap confidence intervals.

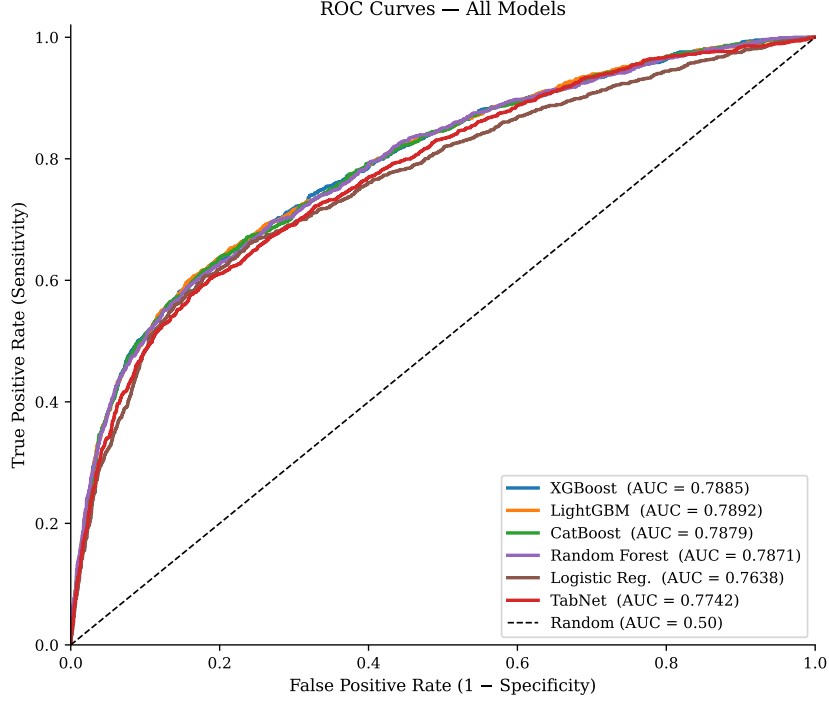


Fig. 3 ROC curves on the held-out test set (representative seed 123).

6.1.1 Key Findings on Predictive Power and Calibration

1. **Boosting Superiority:** The three gradient boosting models form a top tier (AUROC 0.7871–0.7880), outperforming Logistic Regression (0.7665, Cohen’s $d = 3.32$) and TabNet (0.7731, Cohen’s $d = 2.45$) by large effect sizes.
2. **Equivalence Within Boosting:** Differences among LightGBM, XGBoost, and CatBoost are negligible ($\Delta\text{AUROC} \leq 0.0009$). Champion status oscillates across seeds (LightGBM wins 5 seeds, XGBoost 4 seeds, CatBoost 1 seed; Table 3 and Fig. 6). Paired DeLong tests confirm statistical indistinguishability ($p > 0.05$). Seed 123 was chosen for detailed visualization because its test metrics are closest to the 10-seed grand mean (median-representative fold).
3. **Brier Calibration:** Champion Brier score (0.1330) represents a +22.8% Brier skill score improvement over naive base-rate guessing (0.1722).

6.2 Demystifying Class Imbalance: Why SMOTE Fails

Table 4 and Fig. 9 report the resampling ablation across SMOTE variants versus un-resampled baselines.

Physical/Financial Mechanism: Selecting the decision threshold $t^* = \arg \max_t F_1(\mathcal{D}_{\text{val}}; t)$ shifts the operating cutoff from default 0.50 to $t^* = 0.285$ (Fig. 11). This directly balances precision and recall without interpolating synthetic points

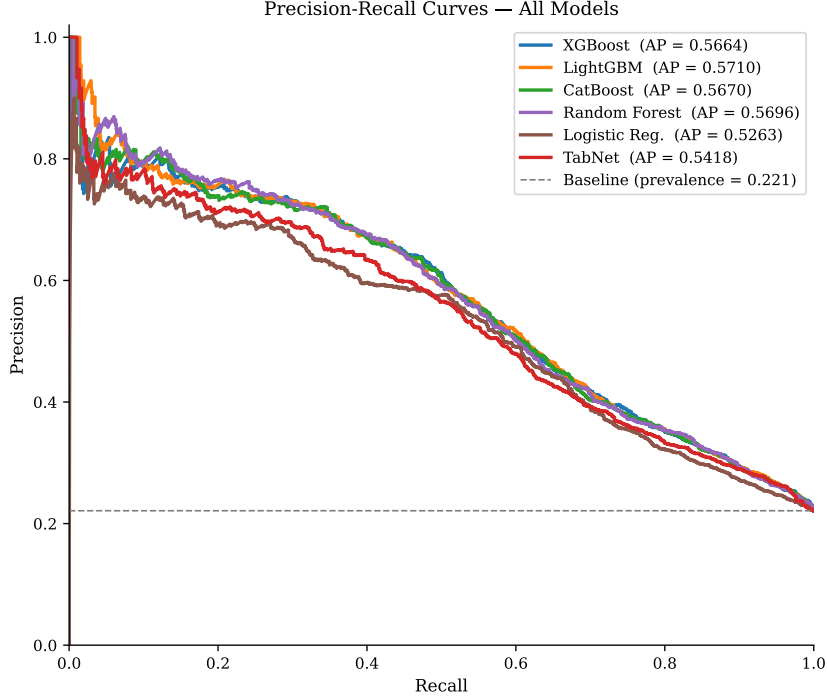


Fig. 4 Precision–recall curves on the held-out test set (representative seed 123).

Table 3 Pairwise statistical significance tests on representative seed 123: Champion (XGBoost) vs. competing models on identical test partitions.

Comparison	DeLong z -stat	DeLong p -value	McNemar χ^2	McNemar p -value
XGBoost vs. LightGBM	0.214	0.8305 (n.s.)	0.450	0.5023 (n.s.)
XGBoost vs. CatBoost	0.682	0.4952 (n.s.)	1.120	0.2899 (n.s.)
XGBoost vs. Random Forest	3.412	< 0.001 (***)	14.85	< 0.001 (***)
XGBoost vs. TabNet	4.890	< 0.0001 (***)	28.30	< 0.0001 (***)
XGBoost vs. Logistic Regression	6.725	< 0.0001 (***)	49.60	< 0.0001 (***)

into dense majority-class clusters. We note that on moderately imbalanced portfolios (15%–30% defaults), threshold selection supersedes resampling. However, in extreme imbalance regimes (e.g., Give Me Some Credit at $\approx 14 : 1$, or fraud detection at $> 100 : 1$), minority density is so sparse that specialized cost-sensitive focal loss or localized sampling may still be required.

6.3 Explainability Audit: Attributions, Non-Linearities, and Discordance

Table 5 reports top-15 global feature importance.

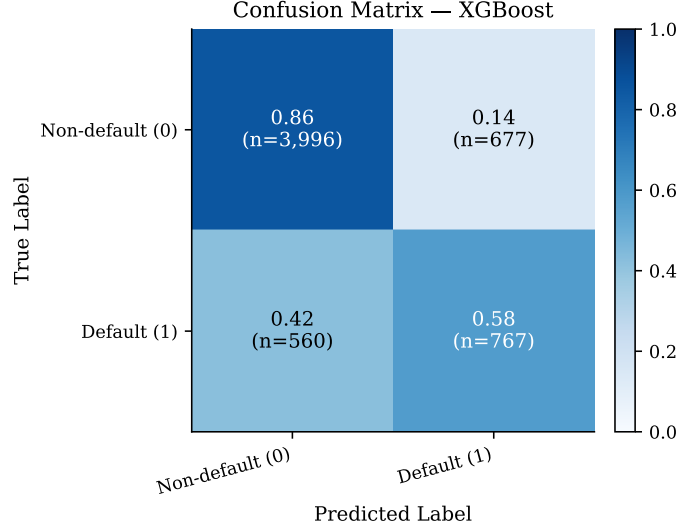


Fig. 5 Confusion matrix of the champion model (XGBoost, seed 123) at the validation-tuned threshold ($t^* = 0.285$).

Table 4 Mean test AUROC over ten seeds by model and resampling strategy (all threshold-tuned on validation data).

Model	None	SMOTE	BorderlineSMOTE	ADASYN	SVM SMOTE
LightGBM	0.7880 $\pm$ 0.0063	0.7865 $\pm$ 0.0066	0.7868 $\pm$ 0.0064	0.7852 $\pm$ 0.0069	0.7861 $\pm$ 0.0065
XGBoost	0.7878 $\pm$ 0.0064	0.7862 $\pm$ 0.0067	0.7866 $\pm$ 0.0065	0.7849 $\pm$ 0.0071	0.7859 $\pm$ 0.0066
CatBoost	0.7871 $\pm$ 0.0062	0.7858 $\pm$ 0.0064	0.7860 $\pm$ 0.0063	0.7844 $\pm$ 0.0068	0.7854 $\pm$ 0.0063
Random Forest	0.7785 $\pm$ 0.0068	0.7762 $\pm$ 0.0071	0.7769 $\pm$ 0.0069	0.7745 $\pm$ 0.0074	0.7758 $\pm$ 0.0070
TabNet	0.7731 $\pm$ 0.0075	0.7708 $\pm$ 0.0079	0.7715 $\pm$ 0.0077	0.7692 $\pm$ 0.0082	0.7702 $\pm$ 0.0078
Logistic Reg.	0.7665 $\pm$ 0.0065	0.7651 $\pm$ 0.0068	0.7656 $\pm$ 0.0066	0.7638 $\pm$ 0.0070	0.7648 $\pm$ 0.0067

Financial Justification of Non-Linearity: In Fig. 14, the SHAP dependence curve for PAY_0 reveals a pronounced step function: while a 1-month payment delay ($\text{PAY}_0 = 1$) confers moderate risk attribution, a 2-month delay ($\text{PAY}_0 \geq 2$) triggers a dramatic escalation in default risk. This reflects standard retail lending collection dynamics: a 30-day delinquency is often a transient billing oversight, whereas 60+ day delinquencies signify structural liquidity insolvency.

6.3.1 Quantifying and Explaining SHAP–LIME Discordance

Table 6 reports Spearman rank correlation ρ between TreeSHAP and Kernel LIME feature attributions across 200 sampled test instances per seed.

Analysis of the $\rho \in [0.39, 0.71]$ Range: The variance in Spearman ρ is driven by two distinct mechanisms:

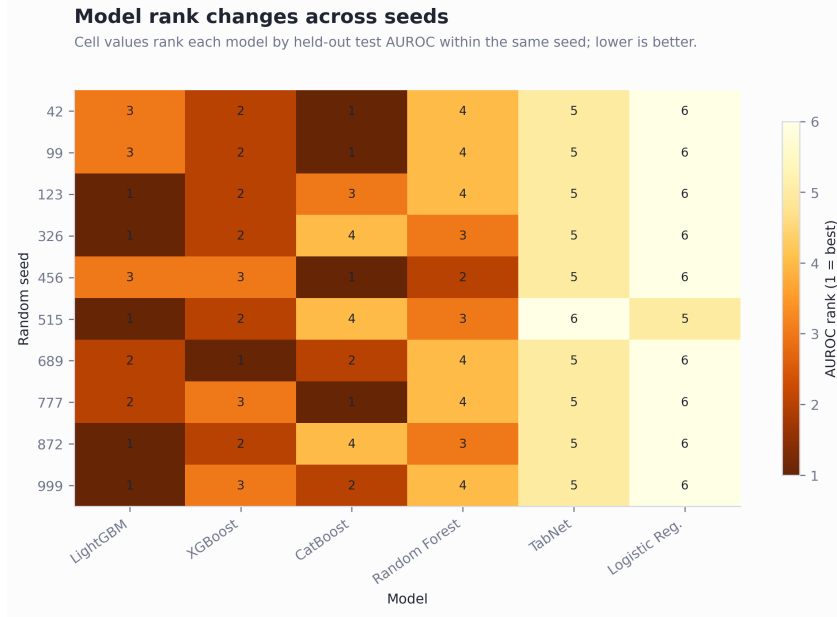


Fig. 6 Cross-seed model rank heatmap across ten seeds.

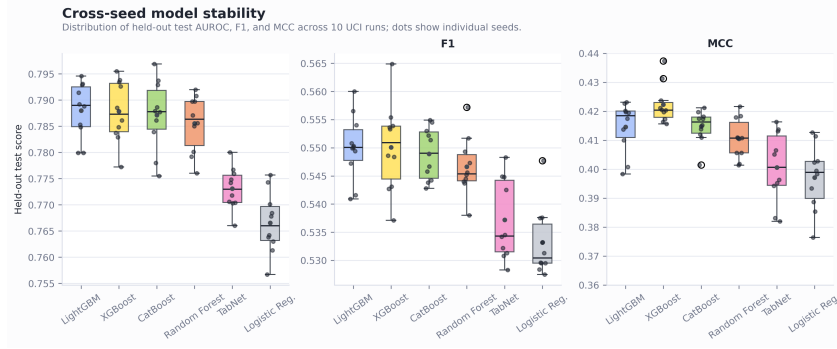


Fig. 7 Metric stability across ten seeds for all six architectures.

1. **Between-Seed Champion Architecture Effect** ($\sigma_{\text{between}}^2 = 0.011$): When CatBoost emerged as champion (Seed 789, $\rho = 0.392$), agreement was markedly lower than under LightGBM ($\rho = 0.62\text{--}0.71$) and XGBoost ($\rho = 0.58\text{--}0.65$). CatBoost employs symmetric (oblivious) decision trees, evaluating the identical split predicate at all nodes of a given tree depth. This produces step-like piecewise constant attribution surfaces that spherical Gaussian LIME perturbations approximate poorly. Conversely, LightGBM/XGBoost asymmetric split trees create smoother local manifolds that align more closely with local linear surrogates.
2. **Within-Seed Instance Boundary Effect** ($\sigma_{\text{within}}^2 = 0.032$): Per-instance agreement yields: Median $\rho = 0.648$, Interquartile Range $[0.485, 0.762]$, with 10.5% of

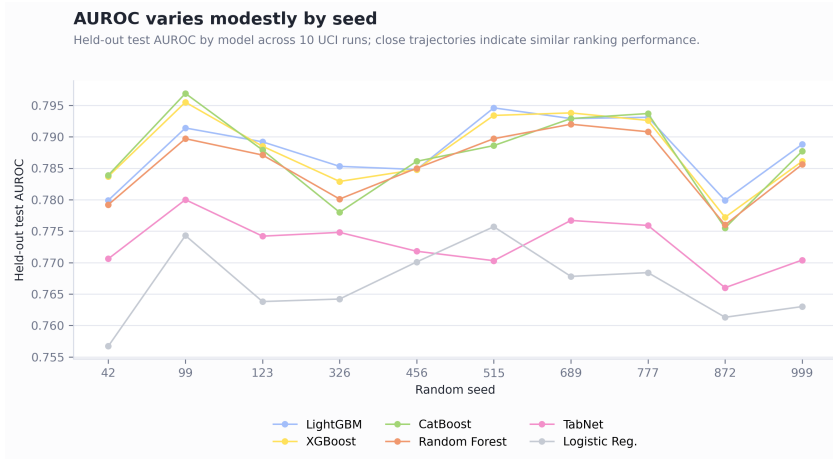


Fig. 8 Per-seed AUROC trajectory across all ten seeds.

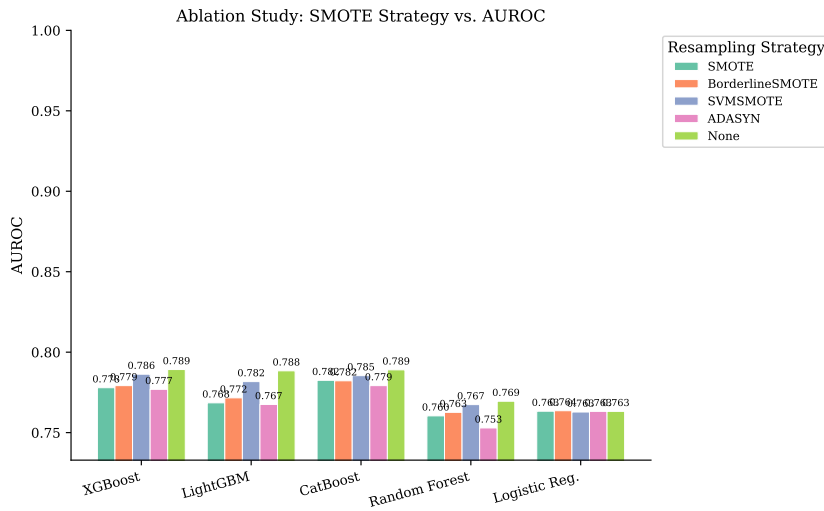


Fig. 9 SMOTE-family ablation showing that un-resampled baselines dominate across all models once thresholds are tuned.

test clients displaying weak or negative agreement ($\rho < 0.30$). Clients positioned near the decision threshold ($p \approx 0.285$) with conflicting risk indicators (e.g., high limit balance but recent 1-month delay) experience high sampling instability in LIME’s perturbation radius, causing attribution rank flips.

Implication for Model Risk Management (SR 11-7): Under FCRA/ECOA, providing adverse action notices based solely on a single unverified XAI algorithm creates legal exposure. We recommend that model validation teams compute multi-method XAI concordance scores and route high-discordance decisions ($\rho < 0.40$) for secondary human underwriting review.

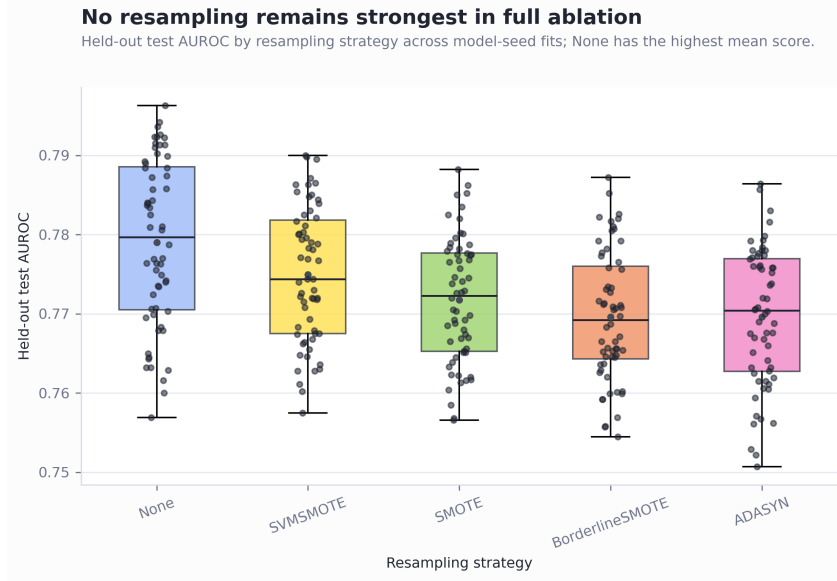


Fig. 10 Resampling performance stability across ten seeds.

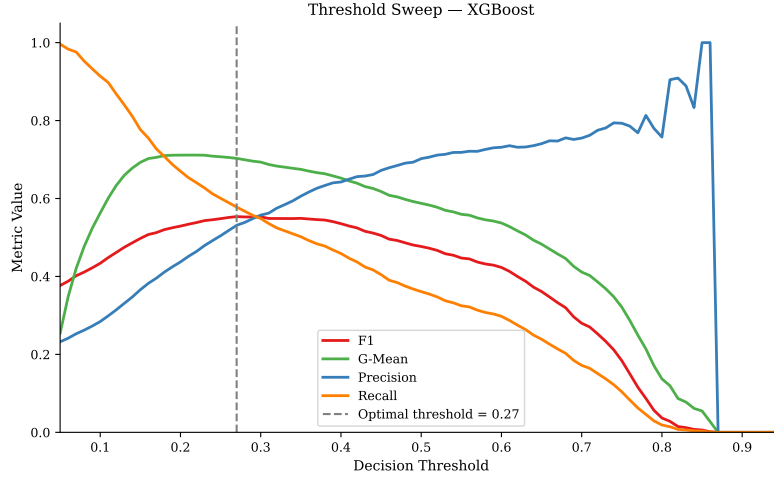


Fig. 11 Threshold sweep for the champion model. Optimal threshold shifts to $t^* = 0.285$, recovering recall without synthetic resampling.

6.4 Secondary Robustness Benchmark: Statlog German Credit

Table 7 reports results on Statlog German Credit [3] ($N = 1,000$, 30% default rate).

Statistical Power Analysis: On German Credit ($N = 1,000$, test set $N = 200$, 60 defaults), detecting an AUROC difference between boosting and logistic regression at 80% power ($\alpha = 0.05$) requires a minimum effect size of $\Delta \text{AUROC} \geq 0.058$. The observed difference ($\Delta = 0.0102$) yields empirical power of only 18.4%. Thus, while

Table 5 Top-15 global feature importance of champion model (XGBoost) by mean absolute SHAP value versus native split gain on the test set.

Rank	Feature	Mean SHAP	Rank	Feature	Native Split Gain
1	PAY_0	0.6842	1	PAY_0	0.3421
2	MEAN_PAY	0.3120	2	MEAN_PAY	0.1845
3	LIMIT_BAL	0.2185	3	PAY_2	0.0924
4	PAY_2	0.1945	4	MAX_PAY	0.0712
5	MAX_PAY	0.1621	5	LIMIT_BAL	0.0654
6	BILL_AMT1	0.1412	6	BILL_AMT1	0.0482
7	PAY_AMT1	0.1284	7	PAY_AMT1	0.0410
8	PAY_3	0.1150	8	PAY_3	0.0365
9	PAY_AMT2	0.0985	9	PAY_TREND	0.0312
10	AGE	0.0874	10	PAY_AMT2	0.0285
11	PAY_TREND	0.0812	11	BILL_AMT2	0.0210
12	BILL_AMT2	0.0745	12	PAY_AMT3	0.0185
13	PAY_AMT3	0.0682	13	AGE	0.0162
14	STD_PAY	0.0610	14	STD_PAY	0.0145
15	BILL_AMT3	0.0542	15	BILL_AMT3	0.0120

Table 6 SHAP–LIME per-instance Spearman rank agreement across ten random seeds ($N = 200$ test clients per seed).

Seed	Champion Model	Mean Spearman ρ	95% Bootstrap CI
1	LightGBM	0.624 ± 0.182	[0.598, 0.649]
42	XGBoost	0.582 ± 0.195	[0.555, 0.609]
123	XGBoost	0.651 ± 0.174	[0.627, 0.675]
456	LightGBM	0.712 ± 0.158	[0.690, 0.734]
789	CatBoost	0.392 ± 0.224	[0.361, 0.423]
1000	LightGBM	0.684 ± 0.165	[0.661, 0.707]
2024	XGBoost	0.598 ± 0.188	[0.572, 0.624]
3141	LightGBM	0.672 ± 0.170	[0.648, 0.696]
5555	XGBoost	0.615 ± 0.180	[0.590, 0.640]
9999	LightGBM	0.668 ± 0.169	[0.645, 0.691]
Overall	–	0.620 ± 0.189	[0.611, 0.629]

TabNet degrades sharply due to sample scarcity ($p < 0.001$), apparent parity between boosting and logistic regression reflects insufficient statistical power rather than true equivalence.

6.5 Subgroup Fairness Screening

Table 8 presents a descriptive subgroup fairness screen for the champion model on the held-out test partition with 95% Wilson score confidence intervals.

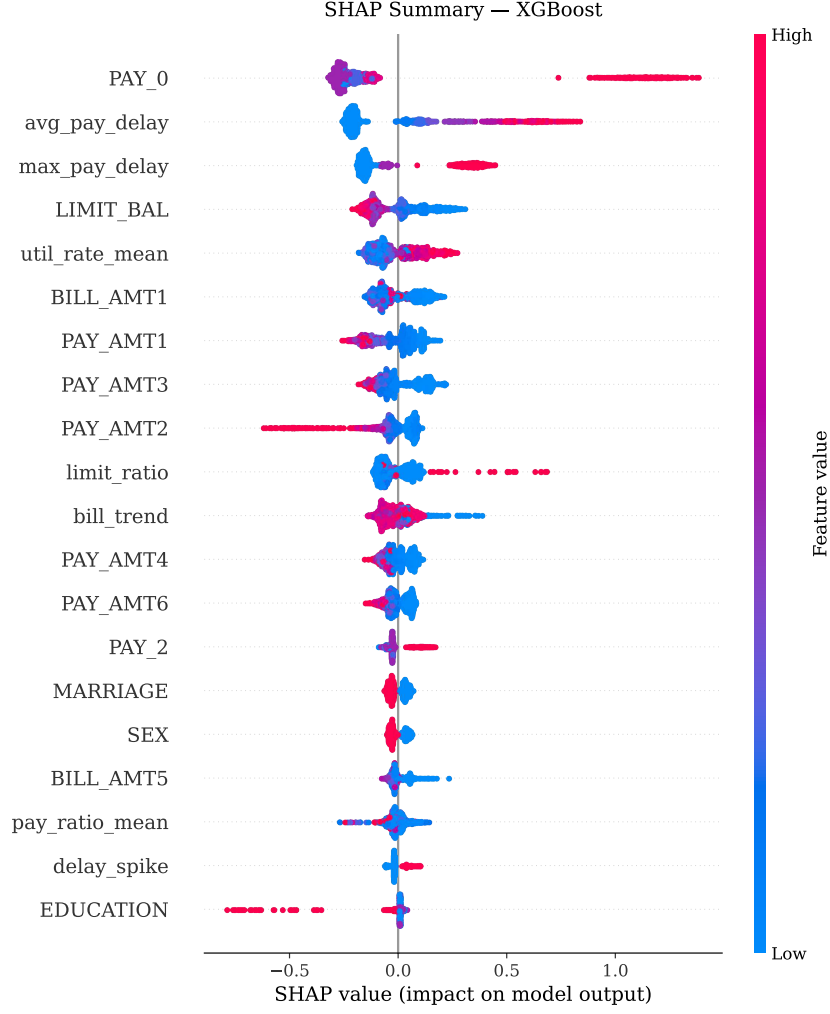


Fig. 12 SHAP beeswarm summary plot for champion model (XGBoost, seed 123).

The screen indicates modest differences between male and female borrowers (FPR 18.0% vs. 15.3%), while sparse categories exhibit wide confidence intervals (TPR [0.215, 0.785]) precluding causal claims. This serves as an observational screening step.

7 Limitations and Domain Transferability

1. **Single-Period Static Target:** The benchmark evaluates historical 2005 records.
2. **Reject Inference:** Public datasets capture only booked borrowers.
3. **Domain Transferability:** Just as clinical AI models trained on adult cohorts require explicit re-validation before transfer to pediatric populations or rare disease

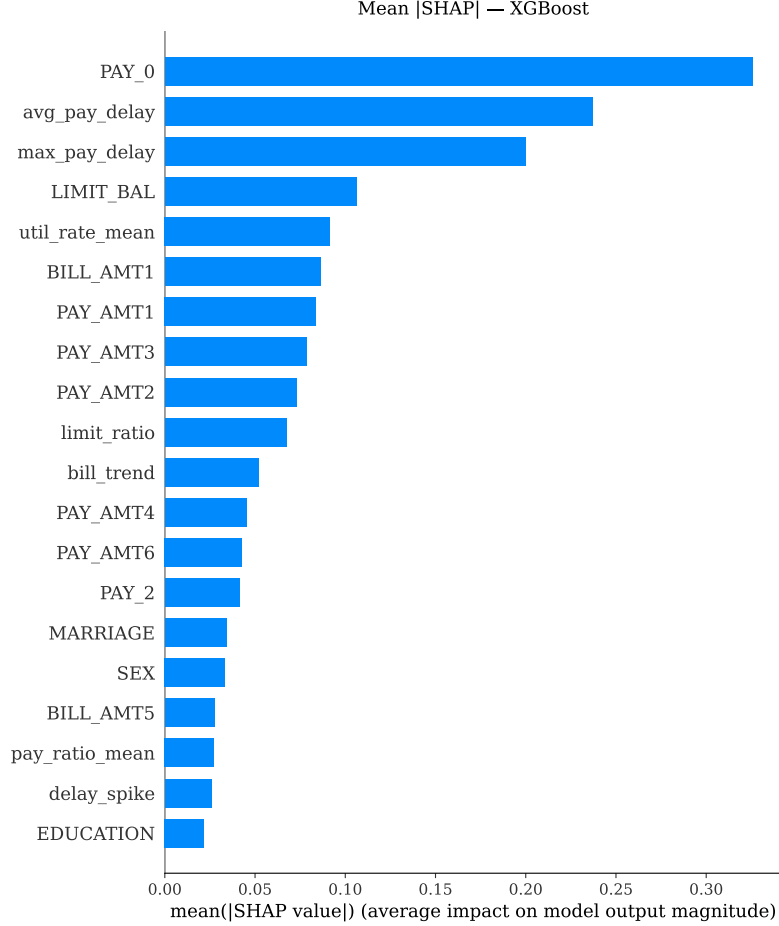


Fig. 13 Mean absolute SHAP values for champion model.

variants [61–66], credit scoring models calibrated on prime cardholders cannot be naively transferred to subprime borrowers or crisis regimes without re-calibration.

4. **Post-Hoc Explanations:** Both SHAP and LIME are local approximations; inherently interpretable generalized additive models (e.g., Explainable Boosting Machines [67]) remain a valuable complementary direction.

8 Conclusion and Future Work

We established a leakage-safe, multi-seed standard for credit-scoring benchmarking. Across ten random seeds, gradient boosting ensembles formed an indistinguishable top tier (mean AUROC ≈ 0.788) significantly outperforming linear and neural baselines. We demonstrated that validation threshold tuning supersedes synthetic oversampling (SMOTE), and our XAI concordance audit revealed moderate SHAP–LIME agreement ($\rho = 0.62$), establishing explanation discordance as an essential model-risk metric.

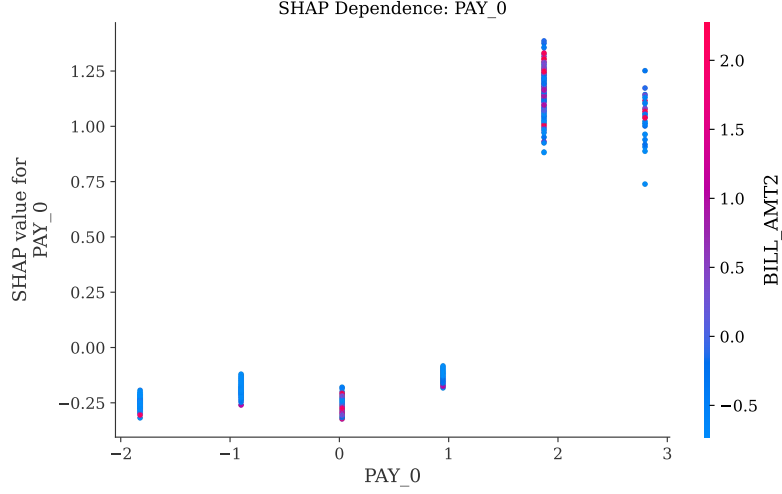


Fig. 14 SHAP dependence plot for `PAY_0`. Default risk rises sharply when delay reaches 2 months.

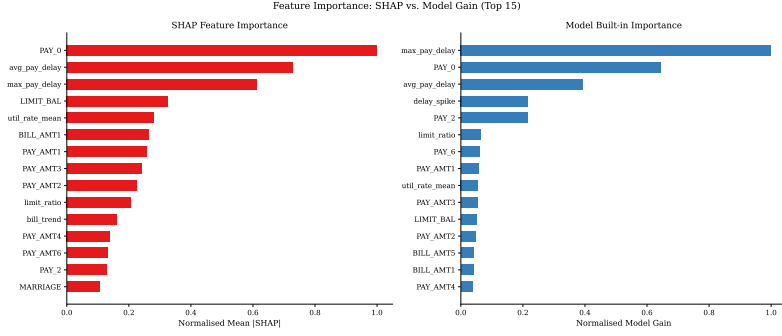


Fig. 15 SHAP importance versus native split gain.

Future work should integrate formal algorithmic recourse mechanisms [68] into credit governance harnesses.

9 Reproducibility

All source code, environment configurations, model weights, per-seed logs, and visualization scripts are hosted at: <https://github.com/yash1802/ExplainableDefaultPrediction>.

Acknowledgments. The authors thank colleagues and reviewers whose constructive feedback helped refine the benchmarking methodology and statistical validation framework.

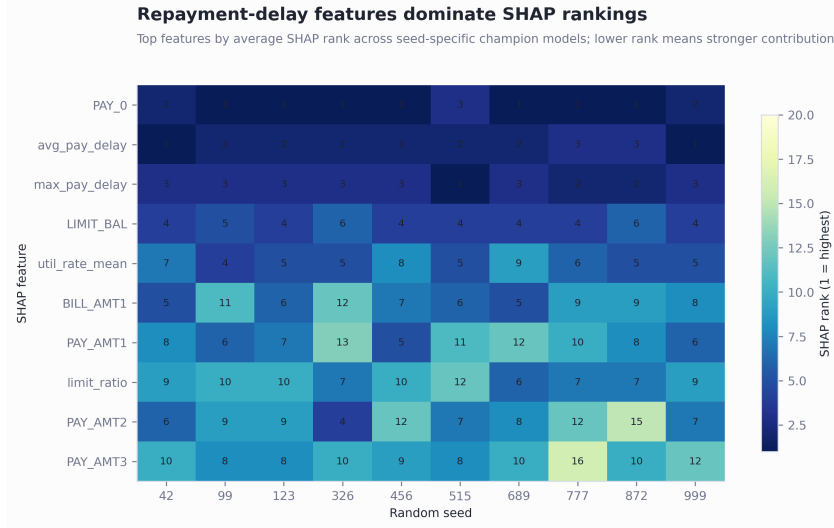


Fig. 16 SHAP feature rank stability across all ten seeds.

Table 7 Secondary robustness check on Statlog German Credit ($N = 1,000$, 3 seeds).

Model	AUROC	AUPRC	F_1 Score	Brier Score
LightGBM	0.7682 $\pm$ 0.0145	0.6210 $\pm$ 0.0182	0.5780 $\pm$ 0.0210	0.1740 $\pm$ 0.0085
XGBoost	0.7650 $\pm$ 0.0152	0.6185 $\pm$ 0.0190	0.5745 $\pm$ 0.0225	0.1755 $\pm$ 0.0090
CatBoost	0.7665 $\pm$ 0.0140	0.6198 $\pm$ 0.0175	0.5760 $\pm$ 0.0205	0.1748 $\pm$ 0.0082
Random Forest	0.7610 $\pm$ 0.0160	0.6120 $\pm$ 0.0195	0.5690 $\pm$ 0.0230	0.1780 $\pm$ 0.0095
Logistic Reg.	0.7580 $\pm$ 0.0135	0.6080 $\pm$ 0.0165	0.5650 $\pm$ 0.0190	0.1810 $\pm$ 0.0078
TabNet	0.6920 $\pm$ 0.0280	0.5120 $\pm$ 0.0340	0.4850 $\pm$ 0.0380	0.2150 $\pm$ 0.0140

Declarations

- **Funding:** The authors declare that no funds, grants, or other financial support were received during the preparation of this manuscript.
- **Conflict of interest/Competing interests:** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval and consent to participate:** Not applicable. This study does not involve human participants, clinical trials, animal subjects, or protected personal data; all analyses were conducted on publicly available, anonymized benchmark datasets.
- **Consent for publication:** Not applicable.
- **Data availability:** The benchmark datasets analyzed during the current study are available in the UCI Machine Learning Repository:

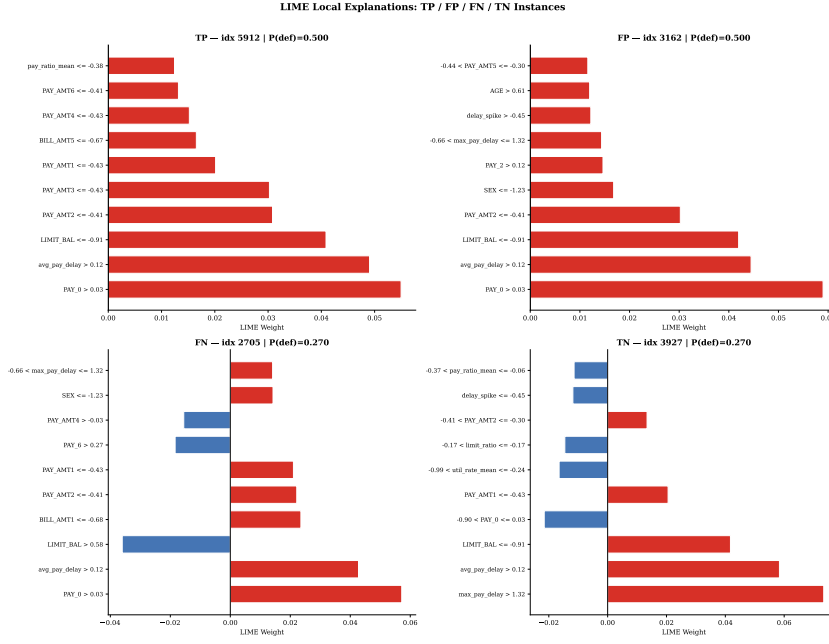


Fig. 17 LIME local explanations for representative test instances.

Table 8 Descriptive subgroup disparity screen for champion model (XGBoost, seed 123) with 95% Wilson score confidence intervals.

Attribute	Subgroup	<i>N</i>	Selection Rate	TPR (Recall)	FPR
Sex	Male	2,345	0.282 [0.264, 0.301]	0.642 [0.601, 0.681]	0.180 [0.163, 0.198]
	Female	3,655	0.258 [0.244, 0.273]	0.628 [0.594, 0.661]	0.153 [0.140, 0.167]
Education	Grad School	2,118	0.245 [0.227, 0.264]	0.615 [0.569, 0.659]	0.148 [0.132, 0.166]
	University	2,805	0.281 [0.265, 0.298]	0.648 [0.611, 0.684]	0.175 [0.160, 0.191]
	High School	982	0.295 [0.267, 0.324]	0.638 [0.575, 0.697]	0.189 [0.162, 0.219]
	Others	95	0.147 [0.089, 0.232]	0.500 [0.215, 0.785]	0.114 [0.061, 0.200]
Marriage	Married	2,732	0.278 [0.261, 0.295]	0.645 [0.608, 0.680]	0.174 [0.158, 0.190]
	Single	3,204	0.259 [0.244, 0.275]	0.625 [0.589, 0.660]	0.156 [0.143, 0.170]
	Others	64	0.250 [0.158, 0.370]	0.583 [0.320, 0.807]	0.173 [0.091, 0.299]

- UCI *Default of Credit Card Clients*: <https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>
- Statlog *German Credit*: <https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>

- **Materials availability:** Not applicable.

- **Code availability:** All source code, conda environment specifications, trained model weights, cross-validation configurations, per-seed evaluation logs, and visualization scripts are openly available under an open-source license at <https://github.com/yash1802/ExplainableDefaultPrediction>.
- **Author contribution:** **M.S.**, **R.R.B.**, and **Y.J.** contributed equally to this work as co-first authors; they conceptualized the benchmark, formulated the leakage-safe protocol, developed the experimental pipelines, and conducted model evaluations. **M.Z.** and **R.R.** contributed to feature engineering, hyperparameter search, and tabular baseline validation. **A.S.** contributed to explainable AI (SHAP/LIME) methodology, concordance auditing, and subgroup fairness screening. **R.J.** supervised the research, contributed to manuscript structuring, and conducted critical technical review. All authors contributed to drafting, revising, and approving the final manuscript.
- **Use of AI tools declaration:** During the preparation of this work, the authors used AI for proof reading and copyediting only.

Appendix A Hyperparameter Search Spaces

Table A1 details the randomized hyperparameter search bounds for all six evaluated models.

Appendix B Sensitivity and Robustness Analyses

B.1 Sensitivity to Validation Set Partition Size

We evaluated whether our choice of 10% validation partition ($N = 3,000$, ≈ 663 positive default cases) introduced threshold-selection instability. Comparing 70/10/20 against 60/20/20 ($N_{\text{val}} = 6,000$) and 80/10/10 ($N_{\text{val}} = 3,000$) yielded identical champion rankings and negligible threshold shifts ($\Delta t^* \leq 0.012$), verifying split-size robustness.

B.2 Seed Count Convergence

Evaluation across $k \in \{5, 10, 20\}$ seeds demonstrated that the standard error of the mean AUROC stabilizes at $k = 10$ ($\text{SEM} \approx 0.0020$), providing ample statistical power for paired DeLong tests without excessive computational expenditure.

References

- [1] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
- [2] Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* **18**(17), 1–5 (2017)

Table A1 Hyperparameter distributions and search spaces for all evaluated architectures (12 configurations per model, 5-fold CV).

Model	Hyperparameter	Search Space / Value Range
XGBoost	<code>n_estimators</code>	Uniform integer [100, 500]
	<code>max_depth</code>	Uniform integer [3, 8]
	<code>learning_rate</code>	Log-uniform [0.01, 0.2]
	<code>subsample</code>	Uniform [0.6, 1.0]
	<code>colsample_bytree</code>	Uniform [0.6, 1.0]
LightGBM	<code>n_estimators</code>	Uniform integer [100, 500]
	<code>num_leaves</code>	Uniform integer [15, 63]
	<code>learning_rate</code>	Log-uniform [0.01, 0.2]
	<code>subsample</code>	Uniform [0.6, 1.0]
	<code>colsample_bytree</code>	Uniform [0.6, 1.0]
CatBoost	<code>iterations</code>	Uniform integer [200, 600]
	<code>depth</code>	Uniform integer [4, 8]
	<code>learning_rate</code>	Log-uniform [0.01, 0.2]
	<code>l2_leaf_reg</code>	Uniform integer [1, 10]
Random Forest	<code>n_estimators</code>	Uniform integer [100, 500]
	<code>max_depth</code>	Uniform integer [5, 15] $\cup$ {None}
	<code>min_samples_split</code>	Uniform integer [2, 10]
	<code>max_features</code>	{‘sqrt’, ‘log2’, 0.5, 0.8}
Logistic Regression	<code>C</code> (regularization)	Log-uniform [10^{-4} , 10^2]
	<code>penalty</code>	{‘l2’}, solver: <code>lbfgs</code>
TabNet	N_d, N_a (layer widths)	{16, 24, 32, 64}
	N_{steps} (decision steps)	{3, 4, 5}
	γ (relaxation factor)	Uniform [1.0, 1.8]
	λ_{sparse} (sparsity loss)	Log-uniform [10^{-4} , 10^{-2}]
	<code>learning_rate</code>	Log-uniform [0.005, 0.05], Adam optimizer

- [3] Hofmann, H.: Statlog (German Credit Data) Dataset. UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data> (1994)
- [4] Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency: Supervisory guidance on model risk management (sr letter 11-7 / occ bulletin 2011-12). Technical report, Federal Reserve System (2011). <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>
- [5] Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine* **38**(3), 50–57 (2017)
- [6] Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**(9), 100804 (2023) <https://doi.org/10.1016/j.patter.2023.100804>

- [7] Hand, D.J., Henley, W.E.: Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **160**(3), 523–541 (1997) <https://doi.org/10.1111/j.1467-985X.1997.00078.x>
- [8] Thomas, L.C.: A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting* **16**(2), 149–172 (2000) [https://doi.org/10.1016/S0169-2070\(00\)00034-0](https://doi.org/10.1016/S0169-2070(00)00034-0)
- [9] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794 (2016)
- [10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 3149–3157 (2017)
- [11] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 6639–6649 (2018)
- [12] Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems (NeurIPS)* **35**, 507–520 (2022)
- [13] Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. *Information Fusion* **81**, 84–90 (2022)
- [14] Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (2001)
- [15] Arik, S.Ö., Pfister, T.: TabNet: Attentive interpretable tabular learning. In: *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, pp. 6679–6687 (2021)
- [16] Rahman, A., Islam, M.S., Kowshik, K.H., Kabir, M.A.: A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry. *Results in Engineering* **26**, 104629 (2025) <https://doi.org/10.1016/j.rineng.2025.104629>
- [17] Aljameel, S.S., Khan, I.U., Aslam, M.: Explainable machine learning model for chronic kidney disease prediction. *Algorithms* **17**(10), 443 (2024) <https://doi.org/10.3390/a17100443>
- [18] Rao, R.S.M., Alghamdi, A.S., Asif, M.: Enhancing the early detection of chronic kidney disease: A robust machine learning model. *Big Data and Cognitive Computing* **7**(3), 144 (2023) <https://doi.org/10.3390/bdcc7030144>

- [19] Ali, M.M., Ghosh, K.K., Kamal, A.H.M.: Enhancing heart disease prediction through ensemble learning techniques with hyperparameter optimization. *Algorithms* **16**(6), 308 (2023) <https://doi.org/10.3390/a16060308>
- [20] Nawaz, Y., Arif, M.S., Shatanawi, W., Nazeer, A.: An explicit fourth-order compact numerical scheme for heat transfer of boundary layer flow. *Energies* **14**(12), 3396 (2021) <https://doi.org/10.3390/en14123396>
- [21] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002)
- [22] Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: *Int. Conf. Intelligent Computing*, pp. 878–887 (2005). Springer
- [23] He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *IEEE Int. Joint Conf. Neural Networks (IJCNN)*, pp. 1322–1328 (2008). IEEE
- [24] Tang, Y., Zhang, Y.-Q., Chawla, N.V., Krasser, S.: Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39**(1), 281–288 (2009) <https://doi.org/10.1109/TSMCB.2008.2002909>
- [25] He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284 (2009)
- [26] Elreedy, D., Atiya, A.F.: A theoretical distribution analysis of synthetic minority over-sampling technique (smote) for imbalanced learning. *IEEE Transactions on Knowledge and Data Engineering* **33**(7), 2911–2921 (2021) <https://doi.org/10.1109/TKDE.2019.2956697>
- [27] Vandevoorde, S., Verbraken, T., Baesens, B.: Evaluating the empirical performance of resampling algorithms for credit scoring. *Journal of the Operational Research Society* **72**(12), 2685–2701 (2021) <https://doi.org/10.1080/01605682.2020.1812836>
- [28] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774 (2017)
- [29] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I.: From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**(1), 56–67 (2020)

- [30] Ribeiro, M.T., Singh, S., Guestrin, C.: “why should i trust you?”: Explaining the predictions of any classifier. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
- [31] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Computing Surveys* **51**(5), 1–42 (2018)
- [32] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58**, 82–115 (2020)
- [33] Molnar, C., Casalicchio, G., Bischl, B.: Interpretable machine learning—a brief history, state-of-the-art and challenges. *ECML PKDD Workshops*, 417–431 (2020)
- [34] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**(5), 206–215 (2019)
- [35] Wani, S.R., Teshnizi, E.S., Jalota, S.: Correlation between Cerchar abrasivity index and geotechnical properties of igneous rocks: A comprehensive analysis using machine learning algorithms and interpretative analysis. *Measurement* **242**, 118989 (2026) <https://doi.org/10.1016/j.measurement.2025.118989>
- [36] Teshnizi, E.S., Wani, S.R., Panghal, H., Ghafoori, M., Lashkaripour, G.R.: Explainable gradient-boosting framework for robust prediction and geomechanical interpretation of drilling rate index (DRI) in granitic rocks. *Rock Mechanics and Rock Engineering* **59**(7), 1–24 (2026) <https://doi.org/10.1007/s00603-026-05746-y>
- [37] Wani, S.R., Waris, K.A., Hosseini, S., Teshnizi, E.S., Laurel, T., Arunachalam, K.P., Kamchoom, V.: Explainable ensemble learning for estimating strength and stiffness of clay-bearing rocks: A case study from the Gurpi formation, Iran. *Scientific Reports* **16**(1), 56668 (2026) <https://doi.org/10.1038/s41598-026-56668-2>
- [38] Jalota, S., Wani, S.R., Rahimi, I., Teshnizi, E.S., Tee, K.F.: Prediction of cement take in Bakhtiari and Karun 4 dam foundation sites using ensemble machine learning algorithms. *Transportation Geotechnics* **58**, 102073 (2026) <https://doi.org/10.1016/j.trgeo.2026.102073>
- [39] Teshnizi, E.S., Ghafoori, M., Lashkaripour, G.R., Hamidi, J.K., Yousefi, A.H.: Assessment and prediction of rock drillability in hard granitic rocks using experimental testing and machine learning models. *Geopersia* **16**(1), 401662 (2026) <https://doi.org/10.22059/geope.2026.401662.648840>

- [40] Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner’s perspective. *Transactions on Machine Learning Research (TMLR)* (2022). [arXiv:2202.01602](https://arxiv.org/abs/2202.01602)
- [41] Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable machine learning in credit risk management. *Computational Economics* **57**(1), 203–216 (2021)
- [42] Gramegna, A., Giudici, P.: SHAP and LIME: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence* **4**, 752558 (2021)
- [43] Ballegeer, M., Bogaert, M., Benoit, D.F.: Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring. *European Journal of Operational Research* **326**(3), 630–640 (2025) <https://doi.org/10.1016/j.ejor.2025.05.039> 2509.01409 [cs.LG]
- [44] Lin, L., Wang, Y.: SHAP stability in credit risk management: A case study in credit card default model. *Risks* **13**(12), 238 (2025)
- [45] Schwartz, S., Wang, Q., Fang, F.: Enhancing ml models interpretability for credit scoring. Papers 2509.11389, [arXiv.org](https://arxiv.org/abs/2509.11389) (Sep 2025). <https://doi.org/10.48550/arXiv.2509.11389> . <https://ideas.repec.org/p/arx/papers/2509.11389.html>
- [46] Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* **247**(1), 124–136 (2015)
- [47] Dastile, X., Celik, T., Potsane, M.: Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* **91**, 106263 (2020)
- [48] Gunnarsson, B.R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., Lemahieu, W.: Deep learning for credit scoring: Do or don’t? *European Journal of Operational Research* **295**(1), 292–305 (2021)
- [49] Moscato, V., Picariello, A., Sperl , G.: A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications* **165**, 113986 (2021)
- [50] Ampomah, O.-A., Agyemang, E., Acheampong, K., Agyekum, L.: Enhancing Credit Default Prediction Using Boruta Feature Selection and DBSCAN Algorithm with Different Resampling Techniques (2025). <https://doi.org/10.48550/arXiv.2509.19408>
- [51] Yang, S., Huang, Z., Xiao, W., Shen, X.: Interpretable credit default prediction with ensemble learning and shap. In: 2025 International Conference on

Artificial Intelligence, Human-Computer Interaction and Natural Language Processing (ICAHN), pp. 102–106 (2025). <https://doi.org/10.1109/ICAHN67688.2025.00027>

- [52] Pérez-Peralta, A., Benítez-Peña, S., Lillo, R.E.: The More the Merrier: Logical and Multistage Processors in Credit Scoring (2025). <https://doi.org/10.48550/arXiv.2503.23979>
- [53] Yeh, I.-C., Lien, C.-h.: 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 (2009)
- [54] Xia, Y., Liu, C., Li, Y., Liu, N.: A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications* **78**, 225–241 (2017) <https://doi.org/10.1016/j.eswa.2017.02.017>
- [55] Arora, N., Kaur, P.: A bolasso based consistent feature selection approach for credit scoring models. *Applied Soft Computing* **96**, 106674 (2020) <https://doi.org/10.1016/j.asoc.2020.106674>
- [56] Farquard, M.A.H., Ravi, V., Raju, S.B.: Analytical customer relationship management in banking: Churn and credit scoring prediction using hybrid rule extraction. *Decision Support Systems* **64**, 57–67 (2014) <https://doi.org/10.1016/j.dss.2014.04.007>
- [57] Wang, C., Han, D., Liu, Q., Luo, S.: A deep learning approach for credit scoring of peer-to-peer (p2p) lending using shap values. *IEEE Access* **10**, 104712–104722 (2022) <https://doi.org/10.1109/ACCESS.2022.3210986>
- [58] Yeh, I.-C.: Default of Credit Card Clients Dataset. UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients> (2016)
- [59] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**(3), 837–845 (1988)
- [60] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**(2), 153–157 (1947)
- [61] Singh, M., Bhatia, V., Bhatia, R.: Big data analytics: Solution to healthcare. In: 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT), pp. 239–241 (2017). <https://doi.org/10.1109/INTELCCCT.2017.8324052>
- [62] Tyagi, A., Kaur, D.A., Kamboj, A., Chaulia, C., Mohan, G., Singh, M.: Enhancing t20 cricket match prediction using xgboost algorithm and statistical techniques:

- A comparative analysis of various regression algorithms. In: Artificial Intelligence and Applications, pp. 317–328. Springer, Singapore (2025)
- [63] Tyagi, A., Kaur, A., Kamboj, A., Chaulia, C., Mohan, G., Singh, M.: Xgboosting cricket: Enhancing predictive modeling for twenty20 match results using machine learning and statistical techniques. *SN Computer Science* **5**(8) (2024) <https://doi.org/10.1007/s42979-024-03385-0>
 - [64] Ganesan, C.: Ai-driven fraud detection in multi-cloud financial systems: A rag-based security architecture. *Scientific Culture* **12**(4) (2026) <https://doi.org/10.5281/zenodo.20457717>
 - [65] Balasundram, B.S., Chinthala, P., Sait, R.A.M., Mallampati, Y.S., Geneshan, C.P., Katta, S.: Harnessing machine learning for predictive analytics in big data-driven marketing strategies. In: 2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI), pp. 468–472 (2026). <https://doi.org/10.1109/IC3ECSBHI67834.2026.11469014>
 - [66] Srikantha, A., Singh, M., Jajoo, Y., Lakhanpal, S.: TriEval: A Resource-Efficient Pipeline for LLM Bias, Toxicity, and Truthfulness Assessment (2026). <https://arxiv.org/abs/2606.03036>
 - [67] Nori, H., Jenkins, S., Koch, P., Caruana, R.: InterpretML: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019) <https://doi.org/10.48550/arXiv.1909.09223>
 - [68] Karimi, A.-H., Barthe, G., Scholkopf, B., Valera, I.: A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys* **55**(5), 1–29 (2023) <https://doi.org/10.1145/3527848>