Ensemble learning for credit scoring: A comparative study of LightGBM and CatBoost

Do Thi Thanh Dieu, Ha Binh Minh, Phan Dinh Phung, and Trinh Hoang Nam
Ho Chi Minh University of Banking

Received 12 February 2026
Revised 02 April 2026
Accepted 12 April 2026

Abstract

Purpose – This paper examines whether gradient-boosted tree ensembles can improve credit-scoring performance over a traditional logistic model while remaining usable within common banking governance practices.

Design/methodology/approach – Using the UCI dataset (30,000 observations), we compare logistic regression, LightGBM, and CatBoost under consistent class weighting ($w = 3.52$). All models are tuned with Optuna (40 trials) using stratified 3-fold cross-validation. Evaluation includes ROC-AUC, PR-AUC, KS, and Brier metrics; DeLong's test; SHAP interpretation; and calibration diagnostics via ECE and SSM regression. Gender fairness and PSI stability checks are also performed.

Findings – Both ensembles outperform logistic regression (ROC-AUC ≈ 0.777 vs 0.710; KS ≈ 0.42 vs 0.36). DeLong's test confirms no significant difference between CatBoost and LightGBM ($p = 0.768$), though both exceed the baseline ($p < 0.001$). SHAP identifies repayment status (PAY_0) and credit limit (LIMIT_BAL) as dominant drivers. Post-hoc isotonic calibration improves LightGBM's ECE from 0.200 to 0.008. Finally, ensembles reduce fairness gaps and maintain high stability (PSI < 0.01).

Originality/value – The paper offers a compact, reproducible evaluation workflow that integrates hyperparameter optimization, imbalance handling, statistical significance testing, explainability, fairness diagnostics, and stability monitoring for ensemble-based credit scoring.

Keywords Calibration, CatBoost, Credit scoring, Ensemble learning, Fairness, LightGBM, SHAP, Stability
Paper type Research paper

1. Introduction

Credit scoring is a foundational component of retail lending, shaping approval decisions, pricing, and portfolio risk control. In many institutions, transparent statistical models-especially logistic regression scorecards-remain popular because they are relatively easy to validate, communicate, and govern, and because they fit long-standing industry practices and oversight expectations (Hand and Henley, 1997).

At the same time, research and practice have increasingly explored machine learning methods to improve risk discrimination, particularly on large, high-dimensional tabular datasets. Benchmarking work in the credit scoring literature has repeatedly shown that modern classification

JEL Classification — G21, G32, C45, C53

 Do Thi Thanh Dieu: Conceptualization, Methodology, Formal analysis, Writing – original draft. Ha Binh Minh: Data curation, Investigation, Writing – review & editing. Phan Dinh Phung: Supervision, Writing – review & editing. Trinh Hoang Nam: Data curation, Investigation, Writing – review & editing. All authors have read and agreed to the published version of the manuscript.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Statement: The authors declare that there are no conflicts of interest regarding the publication of this article.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, Ha Binh Minh, upon reasonable request. The data are not publicly available due to confidentiality considerations.

Declaration of AI-Assisted Technologies: The authors declare that no generative AI or AI-assisted technologies were used in the creation of this manuscript. All content was written, reviewed, and approved solely by the authors.

algorithms can outperform traditional approaches under standard predictive metrics, motivating continued adoption of more flexible models (Lessmann et al., 2015).

Among machine learning approaches, ensemble learning-especially gradient-boosted decision trees-has become a leading choice for structured data due to its ability to capture nonlinearities and feature interactions. LightGBM is designed to improve efficiency and scalability in gradient boosting through algorithmic and engineering optimizations, making it attractive for large-scale model development (Ke et al., 2017). CatBoost further addresses practical modeling challenges by incorporating ordered boosting and dedicated handling of categorical variables to mitigate prediction shift and target leakage concerns (Prokhorenkova et al., 2019). Recent surveys document rapid progress in deep tabular learning (including transformer-based and foundation-model approaches), although gradient-boosted trees remain strong baselines for many structured-data tasks (Badaro et al., 2023; Hollmann et al., 2025; Somvanshi et al., 2024; Ye et al., 2025)the natural language processing community has witnessed advances in neural representations of free texts with transformer-based language models (LMs.

However, performance improvements alone do not guarantee deployability in lending. A key barrier is model transparency: complex ensembles are often perceived as "black boxes" complicating validation, internal model governance, and communication to decision-makers. This is particularly sensitive in high-stakes settings, where the debate about reliance on post-hoc explanations versus inherently interpretable models remains active (Rudin, 2019). Recent scholarship on AI governance in finance shows that emerging risk-based regulatory approaches and responsible-AI governance frameworks increasingly translate these expectations into concrete requirements for transparency, human oversight and lifecycle monitoring (Hulok, 2025; Papagiannidis et al., 2025; Ridzuan et al., 2024).

To improve transparency without sacrificing predictive power, explainable AI (XAI) techniques have become central to applied risk analytics. SHAP provides a unified framework for attributing predictions to feature-level contributions with strong theoretical properties, enabling both global (portfolio-level) and local (case-level) explanations (Lundberg and Lee, 2017). In credit scoring, such explanations can help connect model outputs to domain-relevant drivers (e.g., repayment behavior and credit limits) and can support documentation and review processes.

Beyond interpretability, fairness has become an increasingly important dimension of credit decision modeling. Even when protected attributes are not explicitly used, models may reproduce historical inequities through correlated "proxy" variables or through imbalances in observed outcomes, raising concerns about disparate impact (Barocas and Selbst, 2016). A practical way to operationalize fairness diagnostics in supervised learning is to compare group-level error rates and decision outcomes-such as disparities in true positive rates and false positive rates-consistent with widely discussed fairness definitions like equality of opportunity (Hardt et al., 2016).

A further operational requirement is stability over time. Credit risk models are deployed in environments where borrower behavior, economic conditions, and data pipelines evolve; thus, score distributions may drift even when the model code remains unchanged. In industry practice, the Population Stability Index (PSI) is widely used to flag meaningful distribution shifts, and recent research has examined PSI's statistical properties and interpretation to support better monitoring decisions (Yurdakul and Naranjo, 2020).

Given these considerations, this study proposes an applied and reproducible framework for ensemble-based credit scoring that integrates performance, explainability, fairness, and stability checks in one workflow. Using the UCI Default of Credit Card Clients dataset and its introductory empirical study as a foundational reference (Yeh, 2009; Yeh and Lien, 2009), we benchmark logistic regression against LightGBM and CatBoost. The evaluation combines discrimination and probability quality metrics with SHAP-based explanations, group-level fairness diagnostics (gender-based comparisons), and PSI-based stability monitoring. The combined evidence is intended to support model selection and responsible deployment decisions for credit scoring systems.

2. Literature Review

2.1. Credit scoring: from traditional models to modern benchmarks

Credit scoring has historically been dominated by transparent statistical models, with logistic regression widely adopted due to its simplicity and interpretability (Hand and Henley, 1997). Early empirical comparisons showed that multiple classifiers can be applied to default prediction and that evaluation should rely on ranking and probabilistic measures (Yeh and Lien, 2009). Large-scale benchmarks subsequently reinforced out-of-sample evaluation and often find that

non-linear methods improve AUC/KS relative to linear scorecards, with gains depending on dataset characteristics and imbalance (Lessmann et al., 2015).

2.2. Ensemble learning for credit scoring

Ensemble methods have become a dominant approach for tabular predictive modeling because they capture nonlinearities and interaction effects without requiring extensive manual feature engineering. Random forests (Breiman, 2001) and gradient boosting machines (Friedman, 2001) are two influential families of such methods. In credit scoring, these models often improve discrimination metrics such as ROC-AUC and KS compared with linear baselines, especially when repayment and utilization patterns interact in nonlinear ways (Lessmann et al., 2015). In particular, large-scale empirical benchmarks in the credit-scoring literature often rank tree ensembles-such as random forests and gradient boosting-among the top-performing approaches on AUC/KS-type criteria, while also noting the governance trade-off created by higher model complexity (Lessmann et al., 2015).

Among gradient-boosted tree implementations, LightGBM and CatBoost are frequently used in applied work. LightGBM improves training efficiency via histogram-based splitting and leaf-wise growth, enabling fast learning on large datasets (Ke et al., 2017). CatBoost incorporates ordered boosting and dedicated categorical processing to reduce prediction shift and leakage risks (Prokhorenkova et al., 2019).

Alongside tree ensembles, recent work explores transformer-based architectures and pre-trained/foundation-model approaches for tabular prediction. Surveys published in 2023–2024 summarise techniques such as tabular transformers and attention-based feature representations, while also noting that gradient-boosted trees remain difficult baselines to beat on many real-world datasets (Badaro et al., 2023; Somvanshi et al., 2024). More recently, a tabular foundation model (TabPFN) has been proposed for small-to-medium datasets (Hollmann et al., 2025), and transformer-based scorecard architectures have been explored in credit behavior modelling (Ye et al., 2025).

2.3. Explainable AI in credit risk modeling

The need for explainability is particularly acute in lending, where model outputs must be validated, documented, and communicated to both internal stakeholders and customers. SHAP provides additive feature attributions that support global summaries and local case explanations (Lundberg and Lee, 2017). In applied finance, explanation artefacts can also be incorporated into validation and monitoring workflows (Bussmann et al., 2021).

At the same time, interpretability remains an active research debate. Some authors argue that post-hoc explanations can be insufficient for high-stakes decisions and advocate the use of inherently interpretable models when possible (Rudin, 2019). In practice, many institutions adopt a pragmatic approach: they deploy strong predictive models and supplement them with standardized explanation, validation, and monitoring procedures. This motivates empirical studies that report not only performance but also explainability artifacts that can be incorporated into governance processes.

Recent academic discussions of AI regulation and governance in finance indicate that risk-based regulatory approaches and responsible-AI governance frameworks are converging on requirements for transparency, traceability, human oversight and lifecycle monitoring for high-impact decision systems such as creditworthiness assessment (Hulok, 2025; Papagiannidis et al., 2025; Ridzuan et al., 2024).

2.4. Fairness and group disparities in credit decisions

Fairness concerns arise because historical lending data may reflect structural inequalities or different outcome base rates across groups, and predictive models trained on such data can reproduce or amplify disparities. Legal and policy discussions on algorithmic decision-making highlight how seemingly neutral models can generate disparate impact through correlated predictors (Barocas and Selbst, 2016). In the machine learning literature, group fairness has been formalized through criteria such as equality of opportunity and equalized odds, which focus on differences in error rates across protected groups (Hardt et al., 2016). In credit scoring applications, this often translates into comparing group-level true positive rates, false positive rates, and decision rates under a given operating threshold.

Recent evidence from economics and finance further suggests that improved prediction technology may change who receives credit and at what terms, potentially producing uneven

impacts across borrower groups (Fuster et al., 2022). Therefore, fairness evaluation is increasingly viewed as a necessary complement to accuracy benchmarking. Importantly, fairness assessment does not automatically imply a fairness intervention; many applied studies first report group disparities as diagnostic evidence and then discuss governance options such as threshold adjustments, reweighting, or post-processing policies.

2.5. Model stability and drift monitoring

Credit scoring models operate in dynamic environments, and even stable model code can generate unstable outcomes if data distributions shift. Monitoring score distributions is therefore standard practice. PSI is widely used as a practical drift indicator, and recent research discusses its statistical behaviour and implications for monitoring thresholds (Yurdakul and Naranjo, 2020).

2.6. Summary and research gap

Overall, prior studies establish that ensemble learning methods can improve credit scoring performance, while recent developments in XAI provide tools to increase transparency. However, applied pipelines that jointly report discrimination, probability quality (including compression/calibration), explanation artefacts, group disparities and drift signals-while also harmonising tuning budgets, imbalance handling and statistical significance testing-remain limited. This study addresses that gap with an end-to-end, reproducible workflow centred on LightGBM and CatBoost.

3. Data and Methodology

3.1. Data source and problem setting

This study uses the Default of Credit Card Clients dataset available from the UCI Machine Learning Repository (Yeh, 2009). The dataset was originally introduced for benchmarking probability-of-default (PD) prediction methods in consumer credit, and it remains a widely used reference in the credit scoring literature (Yeh and Lien, 2009). The sample contains 30,000 observations. The target variable indicates whether a client defaults in the subsequent month (binary classification).

The explanatory variables represent common information used in retail credit risk modeling and can be grouped into four categories:

(1) Demographics (e.g., gender/SEX, education/EDUCATION, marital status/MARRIAGE, age/AGE).

(2) Exposure/limit (e.g., credit limit/LIMIT_BAL).

(3) Repayment status history (e.g., PAY_0 and prior months' status variables).

(4) Billing and payment behaviours over recent months (e.g., BILL_AMT* and PAY_AMT*).

These variable groups align with the dataset description and the original empirical analysis (Yeh and Lien, 2009).

Because the dataset includes SEX as a demographic attribute, we additionally perform group-level fairness diagnostics across gender groups. The objective is not to claim causal discrimination, but to quantify whether model errors and decision outcomes differ systematically across groups.

3.2. Train–test split and preprocessing

3.2.1. Train–test split. We split the dataset into training (80%) and test (20%) sets using stratified sampling to preserve the overall default rate in both subsets. This design supports an out-of-sample evaluation of discrimination and probability metrics.

3.2.2. Missing values and imputation. If missing values occur, they are handled via simple imputation:

(1) Numeric variables: median imputation

(2) Categorical variables: most-frequent imputation. This is implemented through standard preprocessing components within the scikit-learn pipeline framework.

3.2.3. Categorical feature handling. We treat SEX, EDUCATION, and MARRIAGE as categorical variables. Two different encoding strategies are used depending on the model family:

(1) Logistic regression and LightGBM use one-hot encoding for categorical variables. This ensures compatibility with linear modeling and also standardizes the representation used for LightGBM in our pipeline.

(2) CatBoost uses native categorical handling, where categorical column indices are passed directly to the training procedure. This leverages CatBoost's algorithmic design for categorical processing (Prokhorenkova et al., 2019).

The dual strategy is chosen to keep the pipeline practical and reproducible: one-hot encoding is robust and widely used, while CatBoost's native handling is one of its documented advantages in tabular tasks with categorical inputs (Prokhorenkova et al., 2019).

To pre-empt concerns about comparability, we emphasise that this design reflects typical practical usage: LightGBM is trained within a scikit-learn pipeline using one-hot encoding to ensure portability and to align preprocessing with the logistic regression baseline, whereas CatBoost is provided with categorical feature indices to leverage its dedicated categorical processing mechanisms. The CatBoost method is explicitly formulated to handle categorical inputs via ordered target statistics and related encodings, and its documentation notes that one-hot encoding is used only as a special case for low-cardinality categories (controlled by the one_hot_max_size setting) (Prokhorenkova et al., 2019).

3.2.4. Class-imbalance handling (cost-sensitive learning). Default events are the minority class in this dataset. To account for class imbalance in a transparent way, we use cost-sensitive learning via class weights derived from the training data ratio w = N_neg/N_pos = 3.52 (18691 non-default vs 5309 default cases).

This weighting is applied consistently to all three models: class_weight for logistic regression and LightGBM, and class_weights for CatBoost. The goal is to improve learning signal for defaults without altering the test-set class distribution.

3.3. Models

We compare three models: a traditional baseline (logistic regression) and two ensemble learning models (LightGBM and CatBoost). The selection reflects common practice in credit scoring studies where a linear scorecard is used as a reference point and advanced machine learning models are assessed for incremental gains (Hand and Henley, 1997; Lessmann et al., 2015).

3.3.1. Logistic regression (baseline model). Rationale: Logistic regression remains a widely used baseline in consumer credit scoring due to interpretability, operational simplicity, and long-standing acceptance in risk governance (Hand and Henley, 1997).

Specification and training: Input representation: numeric variables + one-hot encoded categorical variables (SEX, EDUCATION, MARRIAGE).

Imbalance handling: class weights {0: 1.0, 1: 3.52} (Section 3.2.4).

(1) Hyperparameter tuning: regularization strength (C) and penalty (L1/L2) are tuned with Optuna (40 trials) using stratified three-fold cross-validation to maximise ROC-AUC; the best configuration is C = 0.247 with L1 penalty (Akiba et al., 2019).

(2) Optimization: solver = liblinear; maximum iterations set to 4000 to ensure convergence.

Interpretability: Logistic regression provides coefficient-based interpretation in the transformed feature space (one-hot dimensions). In addition to coefficient interpretation, we later compute SHAP explanations in the encoded space to keep explanation format consistent across models (Lundberg and Lee, 2017).

3.3.2. LightGBM (gradient-boosted tree ensemble). Rationale: Gradient boosting methods often outperform linear models in credit scoring because they can capture nonlinear effects and complex interactions among repayment history, utilization, and payment behaviours (Friedman, 2001; Lessmann et al., 2015). LightGBM is a widely used implementation designed for efficiency and scalability in gradient-boosted decision trees (Ke et al., 2017).

Specification and training: LightGBM hyperparameters are tuned with Optuna (40 trials) under stratified three-fold cross-validation to maximise ROC-AUC. Within each trial, up to 5,000 trees are fitted with early stopping (100 rounds), and the final model is refit on the full training set using the selected best iteration (best_iteration = 242).

(1) Best iteration (early stopping): 242.

(2) Learning rate: 0.0182.

(3) Tree complexity: num_leaves = 255, max_depth = 7, min_child_samples = 30.

(4) Row/feature subsampling: subsample = 0.790, colsample_bytree = 0.667.

(5) Regularization: reg_alpha = 8.007, reg_lambda = 0.130, min_split_gain = 0.315.

(6) Imbalance handling: class weights {0: 1.0, 1: 3.52} (Section 3.2.4).

Feature representation: For comparability with the baseline, LightGBM is trained on the same one-hot encoded feature space as logistic regression. Although LightGBM can handle categorical features using native categorical splits in some settings, we use one-hot encoding to keep preprocessing consistent and reduce implementation variability across environments. All results reported in Section 4 use the tuned configuration above.

Interpretability: SHAP values for LightGBM are computed using a tree explainer on the fitted boosting model (Lundberg and Lee, 2017). Because the model is trained on one-hot features, the SHAP outputs are interpreted as contributions of expanded (encoded) features.

3.3.3. CatBoost (gradient boosting with categorical processing). Rationale: CatBoost is a gradient boosting algorithm specifically designed to handle categorical variables effectively and to reduce prediction shift via ordered boosting, which can be advantageous in real-world tabular datasets (Prokhorenkova et al., 2019). In credit scoring, categorical variables (e.g., demographic or status-coded fields) are common; thus CatBoost is a practical candidate.

Specification and training: CatBoost hyperparameters are tuned with Optuna (40 trials) under stratified three-fold cross-validation to maximise ROC-AUC. Each trial fits up to 5,000 boosting iterations with early stopping (od_wait = 100), and the final model is refit on the full training set using the selected best iteration (best_iteration = 690).

(1) Best iteration (early stopping): 690.

(2) Learning rate: 0.0182.

(3) Tree depth: 10.

(4) Regularization: l2_leaf_reg = 94.64, random_strength = 0.0356.

(5) Stochasticity: bagging_temperature = 0.713, rsm = 0.706.

(6) Objective: Logloss; Evaluation metric: AUC.

(7) Imbalance handling: class_weights = [1.0, 3.52] (Section 3.2.4).

Random seed fixed for reproducibility; verbose logging every 200 iterations.

Feature representation: CatBoost is trained on the raw dataframe with categorical feature indices passed explicitly (i.e., without one-hot encoding). This allows CatBoost to apply its internal categorical processing mechanisms (Prokhorenkova et al., 2019).

Interpretability: SHAP values for CatBoost are computed directly on the raw feature space using a tree explainer. This has an advantage for communication because SHAP feature names remain aligned with original variables (e.g., PAY_0, LIMIT_BAL).

3.4. Model training protocol and reproducibility

All models are developed under a uniform training protocol to ensure a fair comparison. The dataset is first split into stratified training (80%) and test (20%) sets. Hyperparameters are optimised on the training split only using Optuna with the same budget for all models (40 trials) under stratified three-fold cross-validation. For LightGBM and CatBoost, early stopping (100 rounds) is used within each trial to select best_iteration; the final model is then refit on the full training set with the selected hyperparameters and evaluated once on the untouched test set.

The complete Optuna-selected hyperparameter configurations for all three models are provided in Appendix A (Table A1) to support reproducibility.

To support reproducibility, randomness is controlled via fixed seeds and all artifacts are persisted, including trained models and tuning logs (best_params_*.json) as well as run metadata (run_info.json).

3.5. Evaluation metrics

3.5.1. Discrimination metrics. We report:

(1) ROC-AUC, a threshold-independent measure of ranking performance.

(2) PR-AUC, informative under class imbalance because it emphasizes precision–recall trade-offs (Davis and Goadrich, 2006).

(3) KS statistic, frequently used in credit scoring practice as the maximum separation between cumulative distributions of scores for events and non-events.

DeLong's test is used for pairwise statistical comparison of correlated ROC-AUC estimates on the same test set (DeLong et al., 1988).

3.5.2. Probability quality and calibration. Probability-level quality is assessed using: Brier score, which measures the mean squared error of predicted probabilities (Brier, 1950).

Expected calibration error (ECE), summarising the average gap between mean predicted probability and empirical default rate across bins (Guo et al., 2017).

To complement Brier score and standard calibration curves, we conduct a Sorting–Smoothing Method (SSM) check. Predictions are sorted and the observed default rate is smoothed using a moving window, creating a monotone empirical reference curve. A linear regression between predicted probabilities and smoothed observed rates summarises reliability via slope, intercept and R^2; slopes below one indicate probability compression. In addition, we report post-hoc calibrated probabilities for logistic regression and LightGBM using Platt scaling (sigmoid) and isotonic regression with cross-validated fitting on the training split (Platt, 2000; Zadrozny and Elkan, 2002).

3.6. Explainability using SHAP

We apply SHAP to explain model outputs at two levels (Lundberg and Lee, 2017; Molnar, 2025):

(1) Global explainability: ranking features by mean absolute SHAP values and visualizing their distribution through summary plots.

(2) Local explainability: inspecting individual predictions to understand which variables increase or decrease default risk.

For CatBoost, explanations are produced in the original feature space. For LightGBM and logistic regression, explanations are produced in the one-hot feature space consistent with their trained representations. SHAP values are computed on a sample of the test set to ensure computational feasibility.

3.7. Fairness diagnostics

Fairness is assessed across gender groups using group-wise:

(1) AUC (ranking parity), and threshold-dependent metrics.

(2) TPR, FPR, and selection rate.

These metrics relate to widely discussed fairness criteria based on error-rate disparities, including equality of opportunity concepts (Hardt et al., 2016). Because fairness metrics depend on the decision threshold, we compute them at each model's KS-optimal threshold, which reflects an operational point commonly used in credit scoring. Reported disparities are treated as diagnostic signals rather than fairness-optimized outcomes.

3.8. Stability monitoring via PSI

We evaluate score stability using the Population Stability Index (PSI), a standard indicator for detecting changes in score distributions between a reference and a current. Because the dataset lacks a time index, we implement a pseudo-temporal split of the test set into two partitions and compute PSI on score distributions. Quantile-based binning is used to define stable reference bins. Recent work has also examined PSI's statistical properties to support monitoring interpretation (Yurdakul and Naranjo, 2020).

4. Results

4.1. Overall predictive performance

Using the Optuna-tuned, class-weighted models (w = 3.52), Table 1a summarizes test-set performance using ROC-AUC, PR-AUC, KS and Brier score. LightGBM and CatBoost achieve nearly identical discrimination (ROC-AUC = 0.778 and 0.777; KS = 0.423 and 0.427), both outperforming logistic regression (ROC-AUC = 0.710; KS = 0.361). PR-AUC shows the same pattern (0.55–0.56 versus 0.49). Because cost-sensitive learning can distort raw probability scaling, probability reliability is examined separately in Section 4.3 using calibration and compression diagnostics.

Table 1b shows that CatBoost and LightGBM are statistically indistinguishable in AUC ($p = 0.768$), while both ensembles significantly outperform logistic regression ($p < 0.001$).

These differences are visualized in Figure 1, which plots the ROC curves for all three models. CatBoost and LightGBM dominate logistic regression across most false-positive-rate ranges, indicating consistently better ranking. Under class imbalance, ROC curves may be complemented by precision–recall analysis; Figure 2 shows that both ensemble models maintain higher precision at comparable recall levels than the baseline, reinforcing the conclusion drawn from Table 1a.

4.2. Operating threshold selection for downstream analysis

Because fairness and some operational decision metrics depend on a classification threshold,

Table 1a. Test-set performance comparison across models (ROC-AUC, PR-AUC, KS and Brier score)

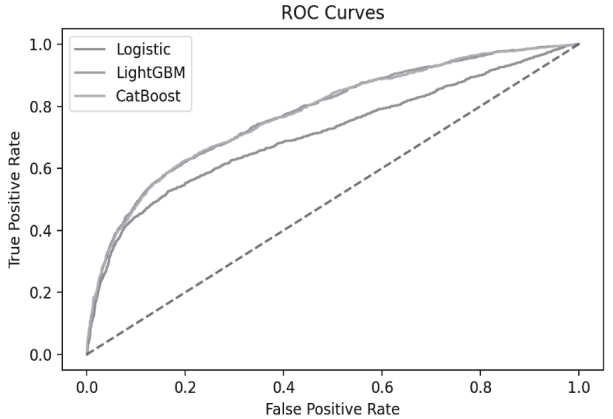
Model	ROC-AUC	PR-AUC	KS	Brier score
CatBoost	0.777	0.556	0.427	0.177
LightGBM	0.778	0.550	0.423	0.179
Logistic regression	0.710	0.491	0.361	0.209

Source(s): Table by the authors

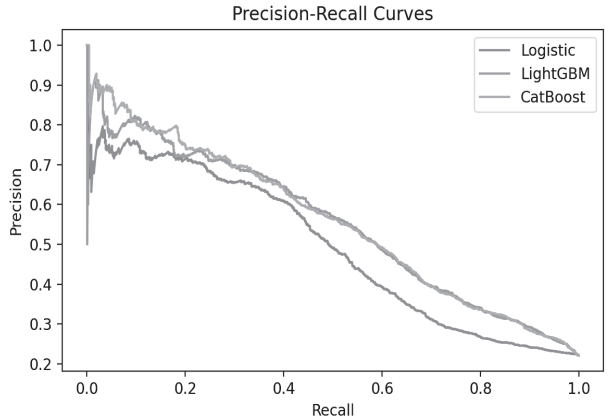
Table 1b. DeLong tests for pairwise ROC-AUC differences on the test set

Comparison	AUC (A)	AUC (B)	p-value (DeLong)
CatBoost vs LightGBM	0.777	0.778	0.768
CatBoost vs Logistic	0.777	0.710	<0.001
LightGBM vs Logistic	0.778	0.710	<0.001

Source(s): Table by the authors



Source(s): Figure by the authors
Figure 1. ROC curves for logistic regression, LightGBM and CatBoost on the test set



Source(s): Figure by the authors
Figure 2. Precision–Recall curves for logistic regression, LightGBM and CatBoost on the test set

we select the KS-maximizing threshold for each model. Table 2 reports thresholds of 0.552 (logistic regression), 0.539 (LightGBM) and 0.496 (CatBoost). Because the models are trained

under cost-sensitive class weighting, these thresholds should be interpreted as operating points for decision making rather than universal probability cut-offs; probability scaling is examined in Section 4.3.

To ensure comparability, the fairness analysis in Section 4.5 is computed using each model's KS-optimal threshold from Table 2.

Table 2. KS-optimal decision thresholds for logistic regression, LightGBM and CatBoost

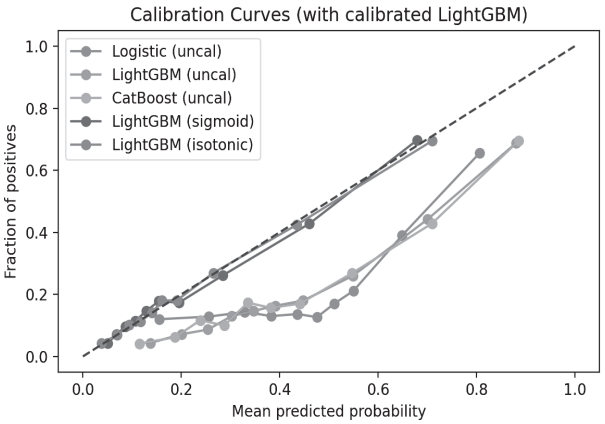
Model	KS-optimal threshold
Logistic regression	0.552
LightGBM	0.539
CatBoost	0.496

Source(s): Table by the authors

4.3. Calibration and probability reliability

In credit scoring applications, ranking ability is critical, but probability quality also matters when predicted probabilities are interpreted as PD estimates. We therefore evaluate probability reliability using both calibration curves and an SSM-based diagnostic.

4.3.1. Calibration curves. Figure 3 presents calibration curves (reliability diagrams) for the three models and illustrates the effect of post-hoc calibration for LightGBM (sigmoid and isotonic). The uncalibrated curves are monotonic but deviate from the diagonal, indicating imperfect probability scaling under class-weighted training. Post-hoc calibration substantially improves LightGBM's reliability, consistent with the quantitative ECE and Brier improvements reported in Table 3b.



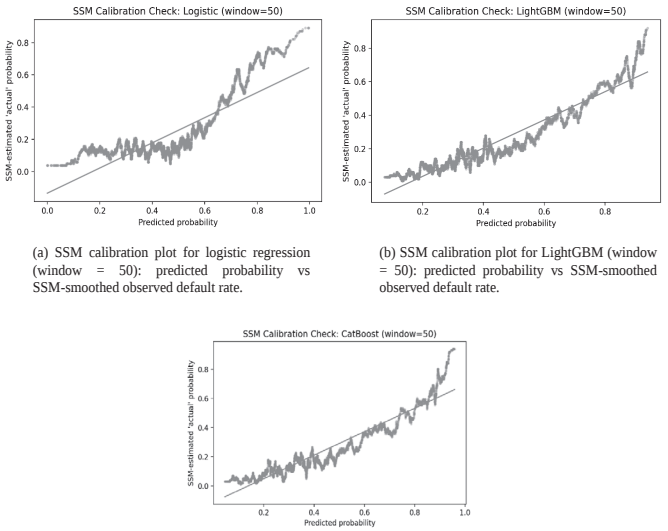
Source(s): Figure by the authors

Figure 3. Calibration curves for logistic regression, LightGBM and CatBoost, including post-hoc calibrated LightGBM (sigmoid and isotonic)

4.3.2. Sorting–Smoothing Method (SSM) calibration check. To provide a complementary calibration view, we apply a Sorting–Smoothing Method (SSM) diagnostic: observations are sorted by predicted probability, the observed default rate is smoothed with a moving window, and a linear regression is fitted between predicted probabilities and the smoothed "empirical" default rate. The results are shown visually in Figures 4a–4c, with a quantitative summary in Table 3a.

Table 3b shows that post-hoc calibration markedly improves probability reliability for LightGBM and logistic regression. For example, LightGBM's ECE decreases from 0.200 (uncalibrated) to 0.008 with isotonic calibration, alongside a lower Brier score (0.135). Logistic regression also improves under sigmoid calibration (ECE = 0.050; Brier = 0.147), providing an additional option for probability refinement when calibrated PD levels are required.

Table 3a highlights probability scaling differences under class-weighted training. Logistic



(a) SSM calibration plot for logistic regression (window = 50): predicted probability vs SSM-smoothed observed default rate.

(b) SSM calibration plot for LightGBM (window = 50): predicted probability vs SSM-smoothed observed default rate.

(c) SSM calibration plot for CatBoost (window = 50): predicted probability vs SSM-smoothed observed default rate.

Source(s): Figure by the authors
Figure 4. SSM-based calibration diagnostics for the three models (window = 50)

Table 3a. SSM calibration regression summary for each model (slope, intercept and R^2)

Model	window	slope	intercept	R^2
Logistic regression	50	0.780	-0.134	0.667
LightGBM	50	0.842	-0.133	0.903
CatBoost	50	0.807	-0.112	0.893

Source(s): Table by the authors

Table 3b. Calibration and compression metrics (Brier and ECE) for uncalibrated and calibrated variants

Model	Variant	Brier	ECE
Logistic	uncalibrated	0.209	0.234
Logistic	sigmoid	0.147	0.050
Logistic	isotonic	0.144	0.011
LightGBM	uncalibrated	0.179	0.200
LightGBM	sigmoid	0.136	0.017
LightGBM	isotonic	0.135	0.008
CatBoost	uncalibrated	0.177	0.192

Source(s): Table by the authors

regression shows a slope well below one (slope = 0.780; intercept = -0.134) and a weaker fit (R^2 = 0.667), indicating substantial compression and noise in the raw probability scale. LightGBM and CatBoost achieve stronger monotonic alignment ($R^2 \approx 0.90$) but still exhibit slopes below one (0.842 and 0.807), suggesting that boosted ensembles can rank risk well while producing compressed probability values that benefit from post-calibration.

Operationally, these results suggest that boosted ensembles can provide excellent ranking while still requiring optional post-calibration when calibrated PD levels are needed for downstream risk computations (e.g., provisioning or limit setting). In our experiments, post-hoc calibration substantially improves probability reliability for LightGBM and logistic regression (Table 3b; Figure 3).

4.4. Explainability analysis (SHAP)

We report model explanations using SHAP to understand key default risk drivers and to support model governance.

4.4.1. Global feature importance. Figures 5a–5c provide SHAP summary plots for logistic regression, LightGBM, and CatBoost, respectively. Across models, repayment status variables and credit exposure consistently appear as dominant drivers. In particular, repayment status (e.g., PAY_0 in the original feature space) emerges as the strongest signal in the tree-based models, and credit limit (LIMIT_BAL) is repeatedly among the most influential predictors. This consistency across different algorithms strengthens confidence that the identified drivers are not model-specific artefacts but reflect robust signal in the dataset.

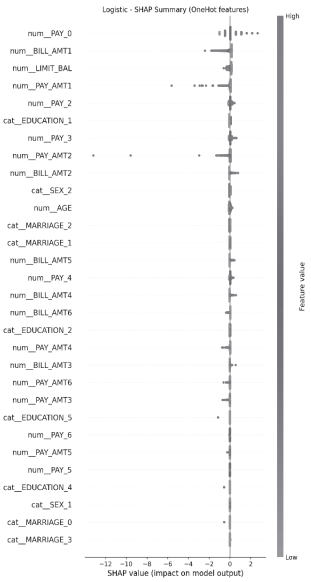
To keep the interpretation concise and comparable across models, Table 4 presents only the ten most influential features for each model, ranked by mean absolute SHAP value. Because logistic regression and LightGBM are trained within a preprocessing pipeline, their SHAP attributions are expressed in the one-hot encoded feature space (e.g., prefixes such as num__ and cat__). In contrast, CatBoost is explained in the original variable space, which preserves the natural meaning of predictors and is generally easier to communicate to non-technical stakeholders. Across all three models, repayment status and recent credit behavior (e.g., PAY_0, payment amounts, and bill amounts) dominate the top ranks, highlighting the central role of delinquency and short-term repayment history in predicting default risk.

Table 4. Top SHAP features ranked by mean absolute SHAP value for each model

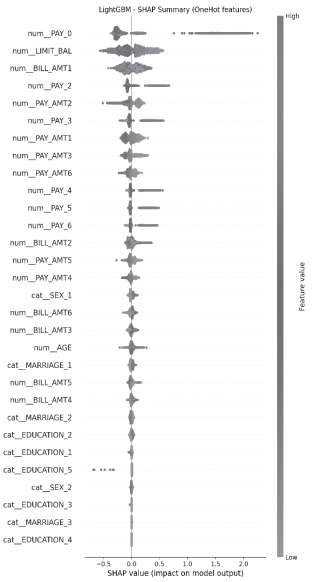
Rank	Logistic (one-hot)	Mean(|SHAP|) Logistic	LightGBM (one-hot)	Mean(|SHAP|) LightGBM	CatBoost (raw)	Mean(|SHAP|) CatBoost
1	num__PAY_0	0.400	num__ PAY_0	0.374	PAY_0	0.375
2	num__BILL_ AMT1	0.197	num__ LIMIT_ BAL	0.207	LIMIT_ BAL	0.265
3	num__LIMIT_ BAL	0.110	num__ BILL_ AMT1	0.127	BILL_ AMT1	0.124
4	num__PAY_ AMT1	0.075	num__ PAY_2	0.125	PAY_ AMT3	0.099
5	num__PAY_2	0.073	num__ PAY_ AMT2	0.099	PAY_ AMT1	0.096
6	cat__ EDUCATION_1	0.072	num__ PAY_3	0.089	PAY_ AMT2	0.088
7	num__PAY_3	0.071	num__ PAY_ AMT1	0.083	PAY_3	0.086
8	num__PAY_ AMT2	0.067	num__ PAY_ AMT3	0.076	PAY_2	0.075
9	num__BILL_ AMT2	0.062	num__ PAY_ AMT6	0.053	PAY_4	0.071
10	cat__SEX_2	0.057	num__ PAY_4	0.051	PAY_6	0.056

Source(s): Table by the authors

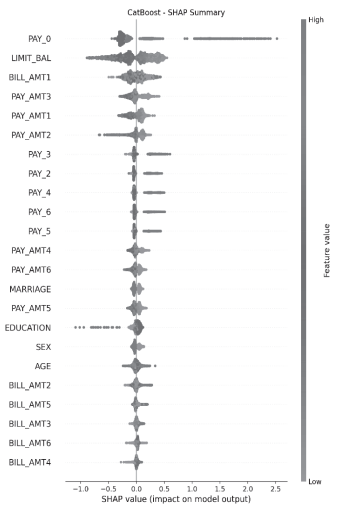
4.4.2. Dependence behavior of the most influential feature. To complement global rankings, Figure 6 shows the SHAP dependence plot for CatBoost's most influential feature. The plot



(a) SHAP summary plot for logistic regression (one-hot feature space).



(b) SHAP summary plot for LightGBM (one-hot feature space).



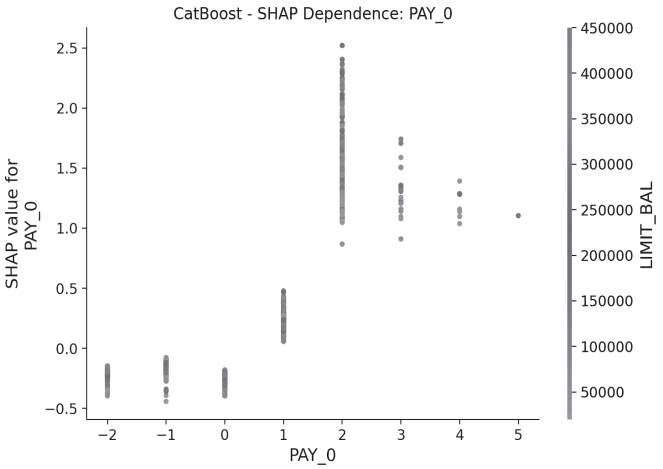
(c) SHAP summary plot for CatBoost (original feature space).

Source(s): Figure by the authors

Figure 5. SHAP summary plots for model explainability across the three models

illustrates how the feature's contribution changes across its value range, providing a nuanced view of non-linear effects. Such dependence diagnostics are valuable in credit scoring governance because they help verify directional plausibility and identify regions where risk contributions change sharply.

Interpretation in credit-risk terms: PAY_0 encodes the most recent repayment status, where higher values indicate delinquency or delay. Accordingly, the dependence pattern implies that moving into delinquent repayment states sharply increases the model's default-risk contribution. LIMIT_BAL represents the granted credit limit (exposure and a proxy for customer creditworthiness); lower limits tend to be associated with higher risk contributions, and the color gradient suggests interaction between repayment status and exposure when explaining risk.



Source(s): Figure by the authors

Figure 6. SHAP dependence plot for CatBoost's most influential feature

4.5. Fairness diagnostics across gender

Fairness is evaluated across gender groups (SEX) using group-level AUC and threshold-dependent metrics: true positive rate (TPR), false positive rate (FPR), and selection rate (the proportion of instances classified as default at the operating threshold). To ensure operational comparability, all results are computed at each model's KS-optimal threshold reported in Table 2.

Tables 5a–5c summarize gender-based metrics for logistic regression, LightGBM and CatBoost at each model's KS-optimal threshold. Group-level AUC values are similar within each model, indicating broadly comparable ranking performance across gender. However, threshold-dependent gaps remain. Logistic regression exhibits the largest TPR gap (0.061), while the ensemble models show smaller gaps (0.026 for LightGBM and 0.029 for CatBoost), suggesting more even identification of defaulters across gender at the selected operating points.

FPR and selection-rate differences are also present. Logistic regression shows the largest FPR gap (0.095) and selection-rate gap (0.095), whereas LightGBM and CatBoost have smaller but non-trivial differences (FPR gaps of 0.048 and 0.051; selection-rate gaps of 0.052 and 0.055, respectively). These results are diagnostic and threshold-specific; they motivate routine reporting of group-level error rates and consideration of monitoring or threshold policies when fairness constraints are operationally required.

4.6. Stability analysis using PSI

We examine score stability using the Population Stability Index (PSI) computed on predicted score

Table 5a. Gender-based fairness metrics for logistic regression at its KS-optimal threshold

Gender	n	AUC	TPR	FPR	Selection rate
Male (1)	2402	0.713	0.560	0.222	0.301
Female (2)	3598	0.705	0.499	0.126	0.206

Source(s): Table by the authors

Table 5b. Gender-based fairness metrics for LightGBM at its KS-optimal threshold

Gender	n	AUC	TPR	FPR	Selection rate
Male (1)	2402	0.773	0.599	0.190	0.286
Female (2)	3598	0.780	0.573	0.142	0.234

Source(s): Table by the authors

Table 5c. Gender-based fairness metrics for CatBoost at its KS-optimal threshold.

Gender	n	AUC	TPR	FPR	Selection rate
Male (1)	2402	0.772	0.636	0.224	0.320
Female (2)	3598	0.779	0.607	0.173	0.265

Source(s): Table by the authors

distributions. Because the dataset lacks an explicit time variable, we perform a pseudo-temporal split of the test set and compare score distributions between two partitions.

Table 6 reports PSI values for each model: logistic regression = 0.0055, LightGBM = 0.0035 and CatBoost = 0.0035. All values are far below common monitoring thresholds (e.g., 0.10), indicating minimal drift under this proxy stability test. While this is not a substitute for time-indexed monitoring in production environments, it provides supportive evidence that the score distributions are not overly sensitive to this partitioning.

Table 6. Score stability measured by population stability index (PSI) under a pseudo-temporal split of the test set

Model	PSI
Logistic regression	0.006
LightGBM	0.003
CatBoost	0.003

Source(s): Table by the authors

4.7. Summary of findings

Across discrimination metrics (Table 1a; Figures 1–2), both ensemble models substantially improve ranking performance over logistic regression. CatBoost and LightGBM are essentially tied in ROC-AUC (0.777) and not statistically different by DeLong's test (Table 1b, $p = 0.768$), while both significantly outperform the baseline ($p < 0.001$). Probability diagnostics (Figure 3; Tables 3 and 3a) indicate that class-weighted training can compress raw probability scales (slopes < 1; higher ECE), and that lightweight post-hoc calibration markedly improves reliability for LightGBM. Explainability results (Figures 5–6; Table 4) consistently identify repayment behavior (PAY_0) and credit exposure (LIMIT_BAL) as primary risk drivers. Fairness diagnostics (Tables 5a–5c) show smaller TPR gaps for ensembles than for logistic regression, although FPR and selection-rate differences persist. Finally, stability monitoring (Table 6) indicates minimal score drift under a pseudo-temporal split (PSI < 0.01).

5. Conclusion

Using the UCI Default of Credit Card Clients benchmark (30,000 observations), this study evaluates whether gradient-boosted tree ensembles improve credit scoring over a traditional logistic model under a governance-oriented and reproducible protocol. To address class imbalance, all models use cost-sensitive class weighting ($w = 3.52$) and are tuned with Optuna under the same budget (40 trials; stratified three-fold cross-validation), with final performance reported on a held-out test set. Both LightGBM and CatBoost materially outperform logistic regression (ROC-AUC ≈ 0.777–0.778 vs 0.710; PR-AUC ≈ 0.55–0.56 vs 0.49; KS ≈ 0.42–0.43 vs 0.36). DeLong's test indicates no statistically significant AUC difference between LightGBM and CatBoost ($p = 0.768$), while both significantly exceed the baseline ($p < 0.001$).

Explainability and probability reliability analyses complement discrimination results. SHAP explanations are consistent across models and align with credit-risk intuition, highlighting recent repayment status (PAY_0) and credit exposure (LIMIT_BAL) as dominant drivers of default risk. However, class-weighted training can yield compressed and miscalibrated uncalibrated probabilities (e.g., LightGBM ECE = 0.200; SSM slopes < 1). Post-hoc calibration substantially improves probability quality (LightGBM isotonic ECE = 0.008; Brier = 0.135; logistic regression also improves under sigmoid/isotonic calibration), supporting safer use of scores when calibrated PD levels are required for downstream risk computations.

From a model-governance perspective, ensembles generally reduce gender-based TPR gaps relative to logistic regression, although threshold-dependent FPR and selection-rate differences remain. PSI values are low (< 0.01) under a proxy split, suggesting limited score drift in this

demonstration setting. A key limitation is that logistic regression and LightGBM explanations are computed in the one-hot encoded feature space, whereas CatBoost explanations remain in the original variable space; an encoding-aligned robustness study is left for future work. Future research should additionally validate these findings on time-indexed credit datasets and explore calibration and fairness interventions for broader protected attributes and operational policies.

Overall, the evidence supports gradient-boosted ensembles as strong candidates for credit scoring when paired with calibration, explainability artifacts, and routine fairness and stability monitoring.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework*. https://doi.org/10.48550/arXiv.1907.10902.

Badaro, G., Saeed, M., & Papotti, P. (2023). Transformers for Tabular Data Representation: A Survey of Models and Applications. *Transactions of the Association for Computational Linguistics*, 11, 227-249. https://doi.org/10.1162/tacl_a_00544.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104, 671-732.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32. https://doi.org/10.1023/A:1010933404324.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. *Comput Econ*, 57, 203-216. https://doi.org/10.1007/s10614-020-10042-0.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240). ACM Press. https://doi.org/10.1145/1143844.1143874.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44, 837-845.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 1189-1232. https://doi.org/10.1214/aos/1013203451.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance*, 77, 5-47.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On Calibration of Modern Neural Networks*. https://doi.org/10.48550/arXiv.1706.04599.

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *J R Stat Soc Ser A Stat Soc*, 160, 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x.

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of Opportunity in Supervised Learning*. https://doi.org/10.48550/arXiv.1610.02413.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637, 319-326. https://doi.org/10.1038/s41586-024-08328-6.

Hulok, M. (2025). The EU model of AI governance: regulating artificial intelligence through law and policy. *ERA Forum*, 26, 527-547. https://doi.org/10.1007/s12027-025-00869-1.

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. In *Advances in Neural Information Processing Systems. Curran Associates*, Inc.

Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247, 124-136. https://doi.org/10.1016/j.ejor.2015.05.030.

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. https://doi.org/10.48550/arXiv.1705.07874.

Molnar, C. (2025). Interpretable Machine Learning: A Guide For Making Black Box Models Explainable. *Christoph Molnar, Munich, Germany*.

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34, 101885. https://doi.org/10.1016/j.jsis.2024.101885.

Platt, J. (2000). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Adv. Large Margin Classif.*, 10.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., & Gulin, A. (2019). *CatBoost: unbiased boosting with categorical features.* https://doi.org/10.48550/arXiv.1706.09516.

Ridzuan, N. N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the Financial Sector: The Line between Innovation, Regulation and Ethical Responsibility. *Information*, 15, 432. https://doi.org/10.3390/info15080432.

Rudin, C. (2019). *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.* https://doi.org/10.48550/arXiv.1811.10154.

Somvanshi, S., Das, S., Javed, S. A., Antariksa, G., & Hossain, A. (2024). *A Survey on Deep Tabular Learning. arXiv.org.* https://arxiv.org/abs/2410.12034v1.

Ye, S., Jiang, Y., Chen, B., & Jiang, C. (2025). Credit Behavior Scorecards Based on TabTransformer and Model Interpretability Exploring. In *Proceedings of the 2025 4th International Conference on Big Data, Information and Computer Network* (pp. 420-428). Association for Computing Machinery. https://doi.org/10.1145/3727353.3727422.

Yeh, I.-C. (2009). Default of credit card clients [Data set]. *UCI Machine Learning Repository.* https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients.

Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36, 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020.

Yurdakul, B., & Naranjo, J. (2020). Statistical properties of the population stability index. *JRMV.* https://doi.org/10.21314/JRMV.2020.227.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 694-699). *Association for Computing Machinery.* https://doi.org/10.1145/775047.775151.

Appendix A. Optuna-selected hyperparameters

This appendix reports the hyperparameter configurations selected via Optuna for each model under stratified 3-fold cross-validation with an equal tuning budget (40 trials per model). Cost-sensitive class weighting was applied using w = N_neg/N_pos = 3.521 (N_pos=5309, N_neg=18691) on the training split.

Table A1. Optuna-selected hyperparameters for logistic regression, LightGBM and CatBoost (class-weighted training)

Model	Tuning (CV; trials)	Class weight	Best CV AUC	Best iteration	Selected hyperparameters
Logistic regression	3-fold; 40 trials	{0: 1.0, 1: 3.521}	0.7274	–	C=0.2474; penalty=l1
LightGBM	3-fold; 40 trials	{0: 1.0, 1: 3.521}	0.7865	242	learning_rate=0.0182; num_leaves=255; max_depth=7; min_child_samples=30; subsample=0.7900; colsample_bytree=0.6674; reg_alpha=8.007; reg_lambda=0.1295; min_split_gain=0.3147
CatBoost	3-fold; 40 trials	[1.0, 3.521]	0.7863	690	learning_rate=0.0182; depth=10; l2_leaf_reg=94.643; random_strength=0.0356; bagging_temperature=0.7130; rsm=0.7059

Source(s): Table by the authors

Corresponding author
Ha Binh Minh can be contacted at: minhhb@hub.edu.vn