Computer and Decision Making

An International Journal

www.comdem.org
eISSN: 3008-1416



Beyond Accuracy: A Multi-Dimensional Explainability and Model-Risk Framework for AI-Based Credit Decisions in Banking

Anamika[1*], Amit Butail[2], Dr. Kritika Rana[3]

[1*]PhD Scholar-Doctor Of Philosophy (CSE), Yogananda School of AI, Computers & Data Science, Institutional Affiliation- Shoolini University, Solan H.P., Orcid Id: (0009-0003-7919-7263), Email ID: anamikasingh@shooliniuniversity.com

[2]PhD Scholar-Doctor Of Philosophy (CSE), Yogananda School of AI, Computers & Data Science , Institutional Affiliation- Shoolini University, Solan H.P., Orcid Id: 0009-0005-0386-4546, Email ID: amitbutail@shooliniuniversity.com

[3]Asst. Professor, Yogananda School of AI, Computers & Data Science, Institutional Affiliation- Shoolini University, Solan H.P., Orcid Id: 0000-0002-2247-5045, Email ID: kritikarana@shooliniuniversity.com

ARTICLE INFO

Article history:
Received 05 April 2026
Received in revised form 30 August 2026
Accepted 11 October 2026
Available online 07 November 2026

Keywords: *mortgage denial prediction; HMDA; explainable artificial intelligence; model-risk .*

ABSTRACT

Beyond predictive discrimination, selecting AI models for mortgage lending requires examining explainability, fairness, calibration, stability, and governance. This study proposes a Beyond-Accuracy, Interpretability and Model-Risk Scoring (BAIMRS) framework and analyzes recorded decisions for 9,360,942 applications in the 2025 Home Mortgage Disclosure Act (HMDA) data. These labels describe historical lender denials, not borrower default; public HMDA data do not include credit scores. Four algorithms (Logistic Regression, Random Forest, XGBoost, and LightGBM) were compared with a monotone LightGBM benchmark on a locked test cohort of 1,211,242 applications and three same-year institution holdouts. LightGBM had the highest ROC-AUC (0.8729) and precision-recall AUC (0.7696), narrowly ahead of XGBoost (0.8727, 0.7688); their paired AUC difference was 0.000179 (95% bootstrap interval -0.000149 to 0.000499). However, LightGBM AUC decreased to 0.7530 at Institution A and its calibration ECE10 reached 0.1129 at Institution C. Following correction of numeric-noise robustness tests, Random Forest and XGBoost had close nominal provisional technical means of 64.362 and 63.209, respectively; XGBoost ranked first under a 0.75 anchor-tolerance scale. Demographic-parity gaps were substantial, and expanded SHAP checks used 300 validation cases per model. Formal deployment rankings remained unavailable owing to incomplete independent governance and auditability evidence. BAIMRS provides a transparent descriptive comparison, not lending approval or regulatory certification.

1. Introduction

Machine learning is now being used to model recorded lender outcomes and rank credit applications for automated credit-decision systems. Recent credit-scoring surveys highlight tensions between predictive quality, explanation and fairness [1] and contemporary surveys document improvements in discrimination and flexibility [2]. The methodological overviews highlight the presence of heterogeneous architectures and evaluation protocols [3] and fintech evidence also shows how nontraditional information can alter the credit assessment process [4]. But a high ROC-AUC does not necessarily imply that the probability estimates are calibrated, explanations are faithful, and outcomes are equitable across subgroups, or operational controls are reproducible, Therefore, AI risk measurement needs to be based on criteria other than a single loss function [5]. The credit market is a good example of how innovations in prediction can have differential impacts [6]. The evidence from fintech discrimination also clearly shows that technological automation does not necessarily remove lending gaps [7]. There are algorithmic and human aspects of bias in mortgage studies [8] and general financial analysis points to a conflict between different definitions of fairness and the goals of different

* *Corresponding author.*
E-mail address: anamikasingh@shooliniuniversity.com

stakeholders [9]. Machine learning is now being used to model recorded lender outcomes and rank credit applications for automated credit-decision systems. Recent credit-scoring surveys highlight tensions between predictive quality, explanation and fairness [1] and contemporary surveys document improvements in discrimination and flexibility [2]. The methodological overviews highlight the presence of heterogeneous architectures and evaluation protocols [3] and fintech evidence also shows how nontraditional information can alter the credit assessment process [4]. But a high ROC-AUC does not necessarily imply that the probability estimates are calibrated, explanations are faithful, and outcomes are equitable across subgroups, or operational controls are reproducible.

Therefore, AI risk measurement needs to be based on criteria other than a single loss function [5]. The credit market is a good example of how innovations in prediction can have differential impacts [6]. The evidence from fintech discrimination also clearly shows that technological automation does not necessarily remove lending gaps [7]. There are algorithmic and human aspects of bias in mortgage studies [8] and general financial analysis points to a conflict between different definitions of fairness and the goals of different stakeholders [9].

While there are various AI criteria in existing conformity frameworks [10], banking model-risk scholarship focuses on documented evaluation and governance in addition to benchmark performance [11]. Transparent credit scoring also must have audit trails and explanations that are independent of the credit scoring process [12]. A key issue that remains to be answered is how to compare technical properties that are not homogeneous, and how to avoid that weak evidence of good governance can be offset by a high accuracy of the model. This research tackles the question by conducting a bounded and evidence-informed assessment of past decisions made by the US mortgage market. It differentiates the predictive task from the default-risk assessment, it includes three same-year institution holdouts (but does not establish out-of-time generalization), and it can assess the effect of scoring choices on technical rankings when the underlying model predictions are fixed. Within the available empirical evidence three objectives were met:

- Measure discrimination, calibration, explainability, fairness, repeated fit stability and perturbation robustness for consistent HMDA evaluation cohorts.
- Develop reference anchored, multi-dimensional technical model-risk profiles and explore sensitivity of weighting, normalization and exclusion of dimensions.
- Identify if the evidence of auditability and governance is available or not for formal deployment eligibility, differentiating between verified artifacts and independent assessments which are not available.

1.1. Research questions and empirically testable propositions

The study asks the following questions: Does predictive-discrimination leadership agree with technical model-risk leadership (H1); is there a detectable difference between the LightGBM and XGBoost ROC-AUCs under paired test-cohort resampling (H2); are there differences in model discrimination and calibration across held-out lending institutions (H3); and do research-weight and normalization-anchor changes affect technical ordering (H4); and is there a lack of evidence of governance and auditability that precludes formal BAIMRS-Core ranking (H5)? H1, H3 and H4 are descriptive assessments, H2 is a paired-bootstrap interval assessment and H5 is an eligibility-control assessment, not a statistically tested causal hypothesis. No proposition equates reproduced historical lending decisions with default prediction or legally validated underwriting.

2. Literature Review, Regulatory Context and Research Gap

Credit models are compared through benchmark studies [13] and interactions in decision-trees are extensions of logistic formulations [14]. Explainable credit-risk modelling can be used for inspection [15] and imbalanced-cohort research can be used to identify class effects [16]. Bank credit assessments incorporate explainability [17], while rescaled cluster-then-predict methods examine transparency-performance trade-offs [18]. There are various ways of achieving transparency, such as interpretable ensembles [19] or segmented modelling [20] and layered interpretation [21]. Credit-approval assistance introduces decision-support issues [22] while selective learning implements abstention under uncertainty [23]. Financial XAI surveys suggest that there are inconsistencies in the definitions [24] and SHAP–LIME comparisons indicate method dependence [25]. Beyond feature ranking, explanation studies in the context of banks [26] and visualizations for lending [27] exist. Quantitative explanation criteria need to be explicit constructs [28]. Plausible interventions and systematic XAI evaluation [29, 30] differentiate plausible interventions from persuasive narratives. There are limitations when using counterfactual benchmarks [31] and constrained searches [32]. Local attribution stability is a separate issue [33] and financial XAI has been reviewed [34] and explainable finance has been covered [35] emphasising

the need for validation.

Creditworthiness fairness evaluation has been proposed as a standardized evaluation [36] and hybrid credit models have been combined with interpretation and disparity analysis [37]. Fairness–profit trade-offs can be measured [38] but rating-level parity is dependent on the definitions of the metrics [39]. There have been discussions on the use of explanation methods in algorithmic lending discrimination [40] and discrimination reviews focus on contextual and causal restrictions [41]. Protected group miscalibration is a different question [42] as is fairness-aware training methods [43] and banking fairness audits [44]. The individual and combined messages of systematic fairness testing [45], methodological difficulties [46] and overview taxonomies [47] caution against the use of one disparity score as a legal finding. There are different explanations for this variation across retraining [48], counterfactual stress scenarios [49] and adaptive credit scoring [50] which provide motivation for robustness checks. Trustworthiness evaluation is based on evidence [51], validation [52] and governance structures [53]. New language-model methods introduce unique risks, and do not necessarily confirm traditional credit models [54]. The gap has to do with joint technical criteria, reproducible anchors, sensitivity and noncompensatory eligibility—not another AUC benchmark.

2.1. Comparative synthesis and research gap

Previous research has tended to focus on a specific validation goal — either predictive performance, lending discrimination, interpretability or financial loss — instead of a well-informed deployment-control decision. The representative research contributions by the constructs addressed are compared in Table 1. BAIMRS does not pretend to be the inventor of these measures; what it does claim is their traceable joint presentation, fixed-anchor sensitivity and a separation between a descriptive technical profile and formally authorized use which is not compensatory. Significantly, this experiment of the present type does not yield independent expert-validated governance scores, adverse-action reason validation or an out-of-time replication.

Table 1. Comparative research synthesis and the contribution boundary of BAIMRS

Reference	Central research emphasis	Relevance to BAIMRS	Evidence boundary
Fuster et al. [6]; Bartlett et al. [7]	Unequal effects and discrimination in credit markets	Motivate explicit protected-group screening	No causal discrimination attribution from observed gaps
von Zahn et al. [10]; Bücker et al. [12]	Fairness, explainability, performance and auditability	Informs multi-criterion and traceability design	Does not supply institutional policy weights
Dumitrescu et al. [14]; Chen et al. [16]	Credit-model predictivity and interpretability	Justifies interpretable baseline comparison	Different targets/cohorts from HMDA lender decisions
Gramegna and Giudici [25]; Nauta et al. [30]	Attribution methods and evaluation quality	Motivates separate fidelity and SHAP diagnostics	Attribution is not an adverse-action legal explanation
Kozodoi et al. [38]; Genovesi et al. [36]	Credit-scoring fairness assessment	Motivates multiple group metrics and support reporting	Group gaps alone do not determine liability
Ballegeer et al. [48]; Nikolaidis and Doumpos [50]	Explanation stability and concept drift	Motivates refits, stress and transfer checks	No 2024-to-2025 temporal test in this analysis
Present BAIMRS framework	Five measured technical dimensions; sixth governance dimension proposed; separate gate	Fixed research anchors, ablations and evidence-dependent eligibility	Full BAIMRS-Core cannot be estimated without approved controls

In the U.S. lending landscape, it is important to make a distinction between legal obligations, data-reporting requirements and supervisory guidance. Prohibited-basis discrimination and adverse-action notification is covered by the Equal Credit Opportunity Act and Regulation B, including the communication of specific reasons [55]. The Fair Housing Act does not allow discrimination in residential real-estate related transactions [56]. Regulation C does not authorize an AI model for credit decisions, but rather regulates HMDA data reporting [57].

On 17 April 2026, the interagency model-risk guidance SR 26-2 was published, which outlines risk-proportionate principles for the development, validation, monitoring and governance of covered traditional and non-generative AI models [58]. Generative and agentic AI are not within the scope of that guidance and any such BAIMRS-G controls are research proposals rather than standards certified by SR 26-2. These sources are related to controls and the evidence that is actually available in Table 2.

Table 2. US credit-decision legal and supervisory context mapped to BAIMRS evidence

Source and model scope	Requirement or principle	Mapped BAIMRS control	Evidence in present study
Regulation B; classical [55]	Prohibited-basis constraints on credit evaluation	Predictor exclusion and proxy-risk audit	Protected fields reserved for audit; proxy effects not fully established
Regulation B; classical/GenAI [55]	Specific principal adverse-action reasons	Reason-code validity and explanation fidelity	SHAP/counterfactual outputs measured; reason concordance untested
Fair Housing Act; classical [56]	Housing-credit nondiscrimination	Subgroup fairness and context-specific review	Descriptive gaps measured; no legal liability finding
Regulation C; data [57]	HMDA disclosure and reporting	Dataset provenance and feature/label coding	LAR decision labels and source provenance described
SR 26-2; classical [58]	Risk-proportionate model development and use	Model documentation, limitations and proportional controls	Research profiles available; bank-specific tiers not approved
SR 26-2; classical [58]	Validation, monitoring and effective challenge	Repeated fitting, stress testing, independent review	Technical tests performed; independent expert validation absent
SR 26-2; classical/vendors [58]	Third-party model oversight appropriate to use	Vendor controls, provenance and inventory	No vendor-model study performed
SR 26-2; GenAI [58]	Generative/agentic models outside guidance scope	Proposed G0–G3 grounding, versioning and oversight	Conceptual only; no GenAI effectiveness claims

The mapping is a research traceability device NOT a compliance certification. Truthful attribution and statutory reason adequacy is distinct from local explanation stability; the paper does not state that the reasons for the adverse action of the underlying creditor have been independently reconstructed. The score thresholds provided in this study as examples should not be interpreted as regulatory-control requirements or supervisory recommendations.

3. Proposed BAIMRS Framework

3.1. Architecture and evidence-based decision control

BAIMRS is not a physical feedback controller, but rather a framework of assessment and model selection for control. It consists of a frozen predictive layer, five measured technical dimensions and one unscored governance dimension, research anchored metric transformations, independent policy thresholds and an auditability evidence gate. As shown in Figure 1, favourable technical outcomes are added to a descriptive profile and independent governance evidence is needed to make a formal eligibility determination. No automatic approval is inferred from a high score.

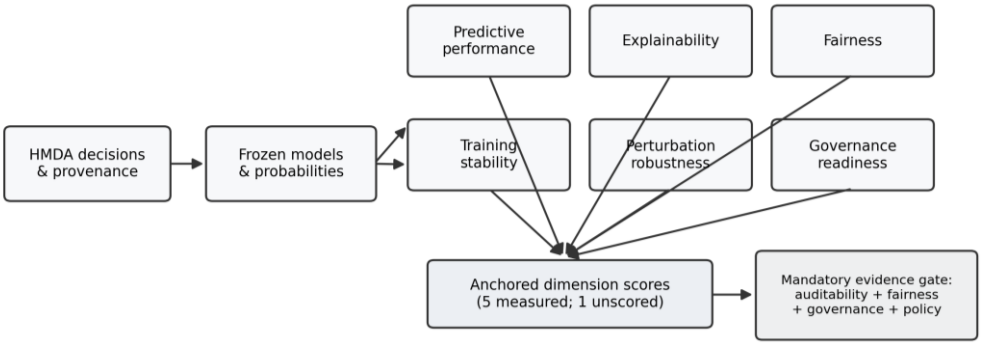


Figure 1. Evidence-based BAIMRS architecture with five measured dimensions and an unscored governance dimension and an independent auditability gate.

3.2. Decision outcome and model output

The binary label for application i is the indicator for the lender denial for that application. Codes 1 and 2 are considered favourable decisions and code 3 is considered denial. Label is not an indicator of nonpayment or borrower default. A model m estimates a denial probability which is converted to a decision by a threshold that is selected only based on validation data:

$$y_i = \begin{cases} 1, & \text{if action_taken}_i = 3 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$\hat{p}_{im} = f_m(x_i), \qquad 0 \le \hat{p}_{im} \le 1 \quad (2)$$

$$\hat{y}_{im} = \mathbb{1}(\hat{p}_{im} \ge \tau_m) \quad (3)$$

3.3. Six scored dimensions and measured inputs

Predictive Performance is a measure of ranking, precision–recall performance, probability quality and threshold dependent discrimination. Explainability is defined as surrogate fidelity, permutation top-five overlap and attribution sparsity; actual SHAP is reported separately. Fairness is an aggregate measure of demographic-parity, equal-opportunity and false-positive-rate disparities, using protected variables that are set aside for auditing purposes. Stability is a measure of the variability in repeated training fits. Robustness is defined as the degradation of AUC/PR-AUC under stress, increase in Brier score and number of decision flips. Independent rubric scores would be needed for Governance Readiness, but are not determined by the current technical experiments

Here, the principle predictive metrics are defined, where the positive class is denial. Balanced accuracy and the fixed-bin expected calibration error are kept as inputs to the evaluation, but are not given causal interpretations:

$$\mathrm{AUC}_m = \Pr(\hat{p}_m^+ > \hat{p}_m^-) + \frac{1}{2}\Pr(\hat{p}_m^+ - \hat{p}_m^-) \quad (4)$$

$$\mathrm{AP}_m = \sum_k (\mathrm{Recall}_k - \mathrm{Recall}_{k-1})\mathrm{Precision}_k \quad (5)$$

$$\mathrm{Brier}_m = \frac{1}{N}\sum_{i=1}^{N} (\hat{p}_{im} - y_i)^2 \quad (6)$$

$$\mathrm{ECE}_{10,m} = \sum_{b=1}^{10} \frac{n_b}{N} |\mathrm{mean}(\hat{p}_{im} \mid b) - \mathrm{mean}(y_i \mid b)| \quad (7)$$

Fairness screening involves comparing the proportions of denial across reported groups g. and the true-denial sensitivity and the false-denial proportions across reported groups g. and the predicted denial. These differences between the highest and lowest scores are descriptive screening measures and do not necessarily indicate illegal treatment.

$$\mathrm{DPgap}_m - \max_g \Pr(\hat{y}_m - 1 \mid g) - \min_g \Pr(\hat{y}_m - 1 \mid g) \quad (8)$$

$$\mathrm{EOgap}_m - \max_g \mathrm{TPR}_{m,g} - \min_g \mathrm{TPR}_{m,g} \quad (9)$$

$$\mathrm{FPRgap_m} = \max_{g} \mathrm{FeR_{m,g}} - \min_{g} \mathrm{FeR_{m,g}} \quad (10)$$

Repeated-training stability is defined as the coefficient of variation of each of the measured performance metrics across refits. Robustness is defined as degradation from a reference input to a specified perturbation condition and decision-flip rate is computed separately.

$$\mathrm{CV_{m,k}} = \frac{\mathrm{SD_r}\left(\mathrm{M_{m,k,r}}\right)}{\left|\mathrm{mean_r}\left(\mathrm{M_{m,k,r}}\right)\right|} \quad (11)$$

$$\Delta \mathrm{M_{m,s}} - \mathrm{M_{m,\,reference}} - \mathrm{M_{m,\,stress(s)}} \quad (12)$$

3.4. Reference anchors and input parameters

Fixed lower and upper research anchors are used for each observed metric (from the implementation, not the candidate-model extrema) to monotonically transform each metric. Clipping will limit the scores shown to 0–100. The nominal 0-0.15 fairness disparity anchors were researcher-selected for illustration rather than elicited or validated as regulatory thresholds. As each model's observed DP gap exceeds 0.15, that component is clipped to zero; a wider DP ceiling of 0.30 is therefore also examined, while EO and FPR ceilings remain unchanged.

$$\mathrm{N_{\uparrow}}(x; L, U) - 100\mathrm{clip}\left(\frac{x - L}{U - L}, 0{,}1\right) \quad (13)$$

$$\mathrm{N_{\downarrow}}(x; L, U) - 100\mathrm{clip}\left(\frac{U - x}{U - L}, 0{,}1\right) \quad (14)$$

Table 3. BAIMRS input metrics, reference anchors and score orientation

Dimension	Metric inputs and research anchor bounds	Preferred direction
Predictive Performance	AUC .50–1; normalized PR gain 0–1; Brier skill 0–1; balanced accuracy .50–1; ECE 0–.10	Higher except ECE
Explainability	Surrogate R^2 0–1; top-five permutation Jaccard 0–1; 90% importance sparsity 0–1	Higher
Fairness	Demographic parity gap 0–.15; EO gap 0–.15; FPR gap 0–.15	Lower
Stability	CV AUC 0–.10; CV PR-AUC 0–.10; CV Brier 0–.10	Lower
Robustness	AUC loss 0–.15; PR-AUC loss 0–.20; Brier increase 0–.05; flip rate 0–.15	Lower
Governance Readiness	Independent governance rubric 0–100; not assessed in the results	Higher; unavailable

Precision–recall gain is normalized with respect to the prevalence of denial and Brier skill is normalized with respect to the constant-prevalence reference. Their transformations keep the direction of improvement and make it impossible to compare raw precision–recall values across cohorts based on prevalence.

$$\mathrm{PRgain_m} = \max\left[0, \frac{\mathrm{AP_m} - \pi}{1 - \pi}\right] \quad (15)$$

Where, $\pi - \Pr(y - 1)$

$$\mathrm{BrierSkill}_{\mathrm{m}} = \max\left[0{,}1 - \frac{\mathrm{Brier}_{\mathrm{m}}}{\pi(1 - \pi)}\right] \quad (16)$$

3.5. Dimension aggregation and noncompensatory eligibility

The implementation uses the arithmetic mean of the metrics that are normalized for each dimension as the available dimension. All declared metrics must be present, no governance evidence is imputed if it is not finite or is not present. The technical profile is an average of the five dimensions measured, and formal BAIMRS-Core

adds in the sixth dimension when verified policy conditions are met.

$$D_{m,d} = \frac{1}{K_d} \sum_{k-1}^{K_d} N_{dk}(x_{m,dk}) \quad (17)$$

$$T_m = \frac{1}{5} \sum_{d \in \{P,X,F,S,R\}} D_{m,d} \quad (18)$$

$$B_m = \frac{1}{6} \sum_{d \in \{P,X,F,S,R,G\}} D_{m,d}, \text{ only if Eligible } _m = 1 \quad (19)$$

$$\text{Eligible}_m = 1\left\{\text{AuditReplay}_m \wedge \text{PolicyApproval}_m \wedge \begin{matrix}\text{FairnessMinimum}_m \\ \wedge\, \text{GovernanceMinimum}_m\end{matrix} \wedge\ \text{CompleteEvidence}_m\right\}$$

An exploratory fairness of 60/100 and a governance minimum of 60/100 is mentioned in the source implementation, neither of which is an independently validated legal limit. The auditability gate also needs to be repeatable and provide evidence for replay and evidence for the assessor besides hashes and split documentation. If a model is ineligible or its eligibility is unknown, B_m is not reported. Weight-draw and single dimension exclusion analyses are not formal deployment suitability tests, but rather tests of stability of T_m. These descriptive sensitivities are reported in Figures 15–17. The optional BAIMRS-G language-model extension is only conceptual and not a proven evaluation.

3.6. Distinct constructs, sensitivity and score interpretation

The six constructs assess different questions: predictive performance is based on observed-outcome discrimination and probability quality; explainability is based on local approximation and attribution characteristics; fairness is based on observed subgroup differences; stability is based on repeated-fitting variability; robustness is based on specified perturbations; and Governance Readiness is based on documented independent controls. Mechanical double counting is avoided by not considering model type as an explanation or governance score. Four model-level dimension scores have only descriptive correlations and are not adequate to demonstrate convergent or discriminant construct validity. The fixed-anchor transformations in (13)–(16) eliminate dependence on the candidate pool, but not the normative decisions about the bounds and the importance of the metric and the subgroups. Equal-weight means are thus research summaries and not institutionally approved suitability thresholds.

3.7. Evidence-gated decision interpretation

The policy gate in (20) is more stringent than the technical gate in (18): the formal composite in (19) cannot be calculated if a reproducible replay is not present, if the governance panel is not present, or if the minimum is not approved. If a control is not found, it is not considered to be evidence of either compliance or regulatory violation, but is considered not verified. The governance rubric and proposed materiality tiers are not calibrated to decisions made by the organization for validation. Weight-sampling and dimension-removal experiments are only descriptive assessments of the technical profile of the policy, and are not a replacement for independent policy approval or legal assessment.

4. Research Methodology

4.1. Research design and data provenance

The evaluation was a retrospective comparison of 2025 HMDA Loan/Application Register data from the official national snapshot [59]. The target was recorded lender denial (action_taken = 3) versus favourable actions (codes 1 and 2), not borrower default or repayment. Public HMDA predictors do not include borrower credit scores. Source integrity, predictor policy, random seeds, and transformed feature inventories were logged. Figure 2 summarizes source validation, isolated model assessment, and evidence-gated reporting.

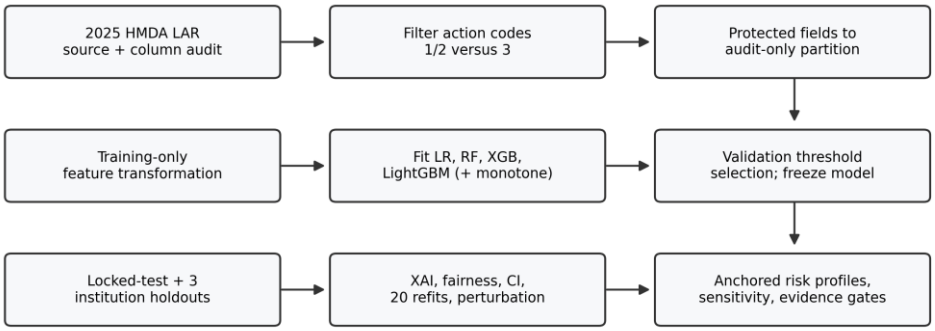


Figure 2. Complete empirical study workflow: HMDA preprocessing, validation-frozen modelling, held-out testing and multidimensional risk evaluation.

4.2. Cohort definition and sampling

Records with action_taken codes 1–3 were retained (N = 9,360,942). The three largest institutions identified by LEI were set aside as institution-level holdouts, after which the remaining observations were deterministically partitioned into training, validation and locked-test cohorts. Table 4 reports the exact cohort counts and denial prevalences. The holdouts originate from the same HMDA year and are not independently sampled external databases.

Table 4. HMDA analytical cohort and institution-level partition characteristics

Cohort	Applications	Recorded denials	Denial prevalence (%)
train	5,712,347	1,300,794	22.77
validation	1,216,995	282,452	23.21
locked_test	1,211,242	278,767	23.01
external_B	501,926	69,523	13.85
external_A	555,321	110,634	19.92
external_C	163,111	72,920	44.71

The reproducibility boundary can be stated without the use of intermediate counts that are not available. Table 5 distinguishes between verified cohort level results and source-field decisions and records that need to be provided by a full extraction log. The analytical total is a post-filter cohort, NOT a count of all the raw 2025 LAR lines. There were no cross-year transfers or independent default-label cohorts.

Table 5. HMDA cohort construction, predictor-use restrictions and evidence provenance

Component	Implemented rule or reported evidence	Analytical interpretation
Source and outcome	2025 HMDA LAR; action_taken 1/2 favourable, 3 denied; 9,360,942 retained	Decision prediction; no observed default target
Partitioning	Training 5,712,347; validation 1,216,995; locked test 1,211,242	Frozen development and held-out assessment
Institution holdouts	A 555,321; B 501,926; C 163,111; same reporting year	Institution transfer; not independent cross-year datasets
Predictor construction	22 raw inputs represented by 27 semantic features; train-only preprocessing	Five numeric and 22 categorical/status predictors
Excluded predictors	action_taken, loan_amount, conforming_loan_limit, LEI, denial reasons, protected attributes	Avoids direct outcome leakage and reserves audit variables
Missingness / special codes	Categorical statuses maintained; numeric missingness imputed from training records	Raw-source filter-by-filter exclusion counts not reproduced in present evidence

Demographic analysis	Race and other protected attributes reserved for auditing; group support recorded	Proxy effects and complete reason-code concordance not established

4.3. Predictor policy and preprocessing

A set of 22 raw HMDA predictor fields was converted to 27 semantic inputs, including five numeric and 22 categorical/status variables (Table 5A). Numeric variables included loan-to-value ratio, loan term, property value, income and affordable-unit counts; debt-to-income ratio was status-coded categorically. Training-fitted transformations imputed numeric missing values, robust-scaled numeric inputs and one-hot encoded categorical inputs; original special-status information was retained. The outcome, loan_amount, conforming_loan_limit, lender identifier, denial reasons and protected audit attributes were excluded from the prediction matrix. Missingness differed across recorded outcomes (Table 5B); prospective availability of underwriting variables at a common decision point was not independently verified. These retrospective predictors therefore cannot rule out decision-time information leakage.

Table 5A. Complete inventory of the 22 raw HMDA predictor fields (27 semantic inputs)

HMDA source	Derived variables	n
loan_to_value_ratio	loan_to_value_ratio_num, loan_to_value_ratio_status	2
loan_term	loan_term_num, loan_term_status	2
property_value	property_value_num, property_value_status	2
multifamily_affordable_units	multifamily_affordable_units_num, multifamily_affordable_units_status	2
income	income_num, income_status	2
preapproval	preapproval_clean	1
loan_type	loan_type_clean	1
loan_purpose	loan_purpose_clean	1
lien_status	lien_status_clean	1
reverse_mortgage	reverse_mortgage_clean	1
open-end_line_of_credit	open_end_line_of_credit_clean	1
business_or_commercial_purpose	business_or_commercial_purpose_clean	1
negative_amortization	negative_amortization_clean	1
interest_only_payment	interest_only_payment_clean	1
balloon_payment	balloon_payment_clean	1
other_nonamortizing_features	other_nonamortizing_features_clean	1
construction_method	construction_method_clean	1
occupancy_type	occupancy_type_clean	1
manufactured_home_secured_property_type	manufactured_home_secured_property_type_clean	1
manufactured_home_land_property_interest	manufactured_home_land_property_interest_clean	1
total_units	total_units_band	1
debt_to_income_ratio	debt_to_income_ratio_public	1

Table 5B. Retrospective source-field missingness by recorded outcome

Raw feature	Favorable (%)	Denied (%)	Denied - favorable (pp)
loan_to_value_ratio	8.87	18.58	9.71
property_value	5.48	12.02	6.54
income	8.58	5.96	-2.62
debt_to_income_ratio	12.42	11.73	-0.69

4.4. Predictive modelling and decision policy

The four major models were Logistic Regression, Random Forest, XGBoost [60] and LightGBM [61]. A monotone constrained LightGBM was used as an additional discrimination benchmark, without an imputed five dimensional profile. Hyperparameters and thresholds were chosen on validation data prior to the frozen locked-test evaluation. Predictions and model configurations were kept separately and the locked-test cohort was not used to tune model thresholds.

4.5. Statistical evaluation and explanation methods

ROC-AUC, average precision, Brier score, expected calibration error in ten bins, specificity, sensitivity, balanced accuracy and F1 were calculated. A comparator that compares a majority class was added. The paired bootstrap uncertainty was based on 1,000 locked-test resamples, which did not incorporate retraining uncertainty for the LightGBM–XGBoost AUC difference. Original model-specific SHAP [62] attributions used 80 observations per model; an independent expanded analysis subsequently evaluated 300 sampled validation observations per model with model-specific additivity checks. Model-agnostic perturbation and surrogate fidelity were presented as separate and LIME [63] was not an experimental result per se, but rather a methodological reference. Bounded counterfactual search noted only the success or failure at reversal, not legal validated actionability.

4.6. Fairness, robustness and generalization

Protected-group audits reported demographic-parity, equal-opportunity and false-positive-rate gaps, with equal-opportunity grounding [64]. All the reported categories were different from the minimum-support sensitivity analysis; ECE and sample size were reported. The variability of the validation-set performance was determined by 20 bootstrap retrainings of the primary models, using replacement samples from training records. Changes in AUC/PR and decision flips were used to assess missingness, numerical-noise and categorical perturbations. Input perturbation stress was evaluated on a validation-only sample; the 15% numeric-noise condition was replayed using frozen models. A separate 20,000-record validation stress sample was used for the matched-threshold comparison and was not pooled with the original stress experiment. Same-year transfer variation was measured by institution-level holdouts.

4.7. BAIMRS evaluation and evidence safeguards

Five observable scores of technical dimensions were created using fixed research anchors. Rank sensitivity was measured by Dirichlet weight draws and anchor perturbation and leave-one-dimension-out ablations. To achieve a formal BAIMRS-Core score, a separate governance rubric was needed along with a weight/anchor policy and evidence of auditability replay. Missing independent human-assessor ratings, policy authorization and replay evidence were not imputed as favorable scores. Model-development and evaluation boundaries were not allowed to change during the scoring exercise.

4.8. Analytical reproducibility and inferential boundaries

In each model, the fitting is done with the training observations and the decision threshold is chosen with the validation cohort; locked-test predictions are done without redoing the threshold. There are 20 resampled training refits to assess the variability of training on a fixed validation sample and 1000 paired bootstrap draws on locked predictions to estimate conditional sampling uncertainty. Robustness replays and matched-threshold stress tests were conducted on separate validation-only diagnostic samples, not the locked test. Model-agnostic surrogate fidelity, attribution sparsity and top-five overlap are not equivalent to native SHAP additivity, and counterfactual numerical validity is not equivalent to legally feasible borrower recourse. For composite scores or unperformed temporal replication or expert-elicited governance weights, no inferential claim is made for missing scores.

5. Experimental Results

5.1. Study population and measured class balance

There were 8140584 applications in the development/locked-test population and 1220358 records in the three institution holdouts. The prevalence of locked-test denials was 23.01% with institution A having 19.92%, institution B 13.85% and institution C 44.71%. This variation led to the need to consider institution-level rather than aggregate-level calibration and selection-rate analysis, as opposed to assuming one aggregate prevalence to be universal.

5.2. Predictive discrimination and class-imbalance performance

Table 6 shows a summary of six classifiers. LightGBM achieved the highest reported ROC-AUC (0.872904), average precision (0.769600), balanced accuracy (0.784609) and lowest Brier score (0.099573). XGBoost was close (AUC 0.872724; AP 0.768762; Brier 0.099785), whereas Random Forest reached AUC 0.862835. Logistic Regression achieved the AUC value of 0.736050 while constrained LightGBM achieved the AUC value of 0.867894. The ROC curves are shown in figure 3 and the precision–recall curves are shown in figure 4. The majority class reference had AUC 0.500 and AP matched the prevalence of denials, 0.230150.

Table 6. Locked-test predictive performance and probability calibration

Model	ROC-AUC	PR-AUC	Brier ↓	ECE10 ↓	Bal. acc.
Logistic Regression	0.7361	0.4528	0.1652	0.0873	0.6763
Random Forest	0.8628	0.7538	0.1048	0.0255	0.7801
XGBoost	0.8727	0.7688	0.0998	0.0086	0.7836
LightGBM	0.8729	0.7696	0.0996	0.0089	0.7846
Monotone LightGBM	0.8679	0.7600	0.1019	0.0091	0.7789
Majority class baseline	0.5000	0.2301	0.2301	0.2301	0.5000

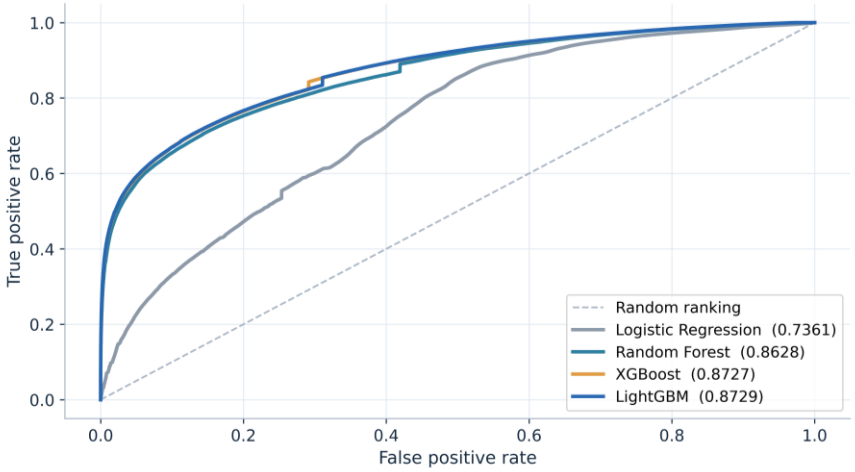


Figure 3. Locked-test receiver operating characteristic curves for recorded HMDA denials.

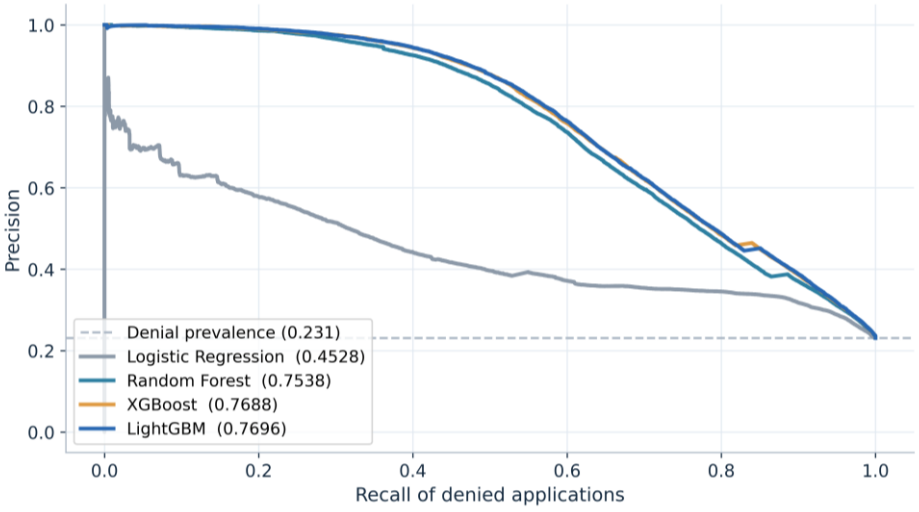


Figure 4. Precision–recall curves with denial treated as the positive class.

5.3. Calibration of predicted denial probabilities

The calibration errors of both the main models (LightGBM and XGBoost) were lower than the other models, with LightGBM having an ECE10 value of 0.008864 and XGBoost with 0.008606. Random Forest achieved ECE10 0.025496 and Logistic Regression came up with 0.087288. The reliability curves are shown in figure 5. These results are for aggregate probability calibration in the locked cohort and do not reflect the calibration or suitability of the subgroups or institutions.

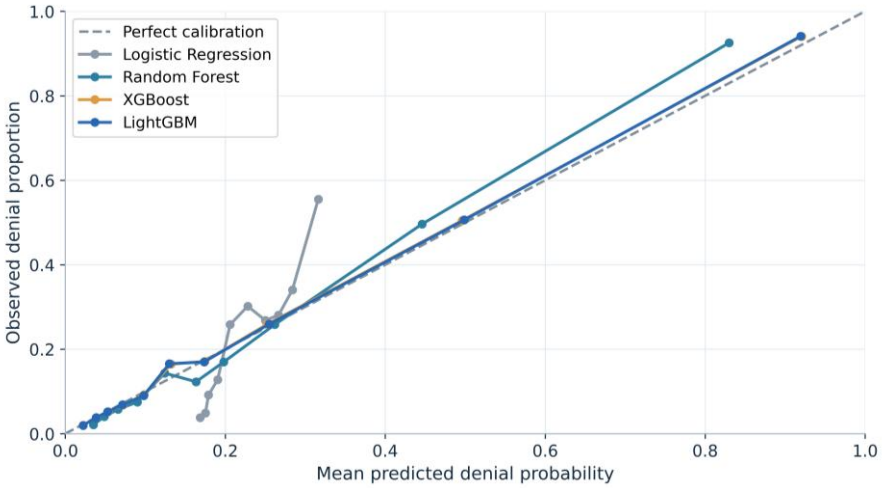


Figure 5. Reliability curves comparing predicted denial probability with observed denial frequency.

5.4. Uncertainty in discrimination differences

The difference between the ROC-AUC of these two models using paired resampling was 0.000179 (95% CI −0.000149, 0.000499), which was considered negligible. The available evidence does not support the hypothesis of LightGBM being able to discriminate better on the interval (including 0). The fraction of paired replicates that favour LightGBM was not a regulatory or deployment decision probability, but rather was 0.85. The paired comparison is shown in Tables 7 and in Figure 6.

Table 7. Paired-bootstrap uncertainty for the LightGBM–XGBoost AUC difference

Comparison	Observed ΔAUC	95% lower	95% upper	Bootstrap draws
LightGBM minus XGBoost ROC-AUC	0.000179	-0.000149	0.000499	1,000

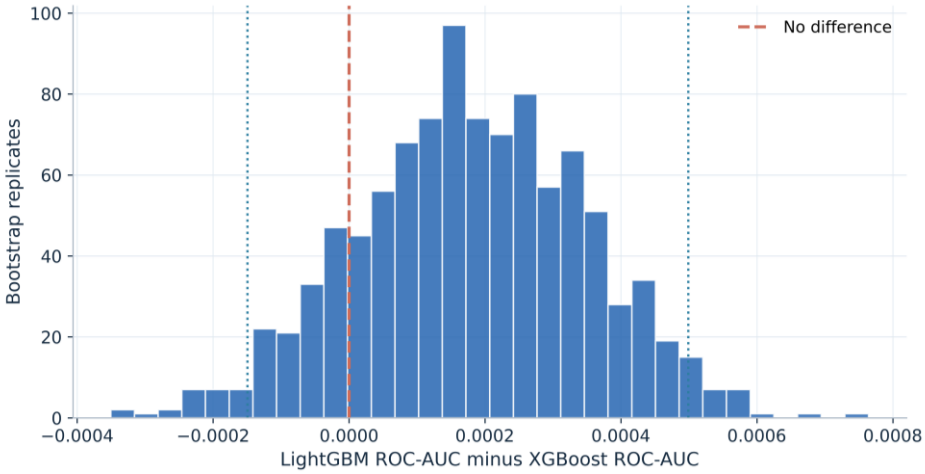


Figure 6. Paired bootstrap distribution of the LightGBM minus XGBoost ROC-AUC difference.

5.5. Operating-point and selection-rate sensitivity

The original validation-frozen denial thresholds were Logistic Regression 0.20, Random Forest 0.24, XGBoost 0.19 and LightGBM 0.19. For a common predicted-denial policy, the LightGBM validation denial rate (32.7077%)

was the reference; matched thresholds were approximately 0.25457, 0.20803, 0.19024 and 0.19000, respectively (Table 7A). The resulting locked-test denial rates were 32.785%, 32.400%, 32.005% and 31.986%. Threshold switching changed 27.217% of Logistic Regression decisions and 5.641% of Random Forest decisions, but fewer than 0.04% for XGBoost and none for LightGBM. Unlike ROC-AUC and PR-AUC, threshold-dependent fairness gaps and decision-flip rates can change at these operating points. Figure 7 compares the original and matched policies; the locked test was not used to select thresholds.

Table 7A. Validation-frozen and matched-selection denial thresholds for all four models

Model	Policy	Threshold	Validation denial rate	Test denial rate	Threshold-switch flip rate
Logistic Regression	Original	0.2000	0.6023	0.6000	0
Logistic Regression	Matched	0.2546	0.3253	0.3279	0.2722
Random Forest	Original	0.2400	0.2709	0.2676	0
Random Forest	Matched	0.2080	0.3271	0.3240	0.0564
XGBoost	Original	0.1900	0.3275	0.3204	0
XGBoost	Matched	0.1902	0.3271	0.3200	0.0004
LightGBM	Original	0.1900	0.3271	0.3199	0
LightGBM	Matched	0.1900	0.3271	0.3199	0

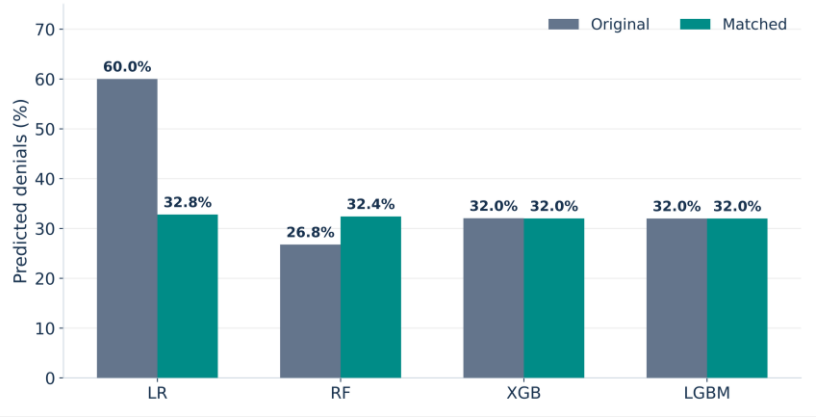


Figure 7. Locked-test denial rates under original and validation-matched thresholds (R8).

5.6. Institution-level generalization

Marked institution heterogeneity is evident in Table 8. LightGBM ROC-AUC was 0.753009 at Institution A, 0.846440 at B and 0.892159 at C, versus 0.872904 on the locked test. These different outcomes may reflect distinct institution policies, applicant mixes and denial prevalences (13.85%-44.71% in the holdouts), although their relative contributions were not identified. At Institution C, LightGBM ECE10 was 0.112934, versus 0.008864 on the locked test; Random Forest had an even higher Institution C ECE10 (0.1299), whereas XGBoost had 0.1080. Figure 8 illustrates differences in discrimination, not calibration. These holdouts are same-year institution transfers, not temporal validation or default prediction.

Table 8. Institution-holdout discrimination and calibration

Institution	Model	Records	ROC-AUC	ECE10
A	Random Forest	555,321	0.7403	0.0421
A	XGBoost	555,321	0.7527	0.0316
A	LightGBM	555,321	0.7530	0.0329

Institution	Model	Records	ROC-AUC	ECE10
B	Random Forest	501,926	0.8349	0.0452
B	XGBoost	501,926	0.8440	0.0372
B	LightGBM	501,926	0.8464	0.0362
C	Random Forest	163,111	0.8732	0.1299
C	XGBoost	163,111	0.8892	0.1080
C	LightGBM	163,111	0.8922	0.1129

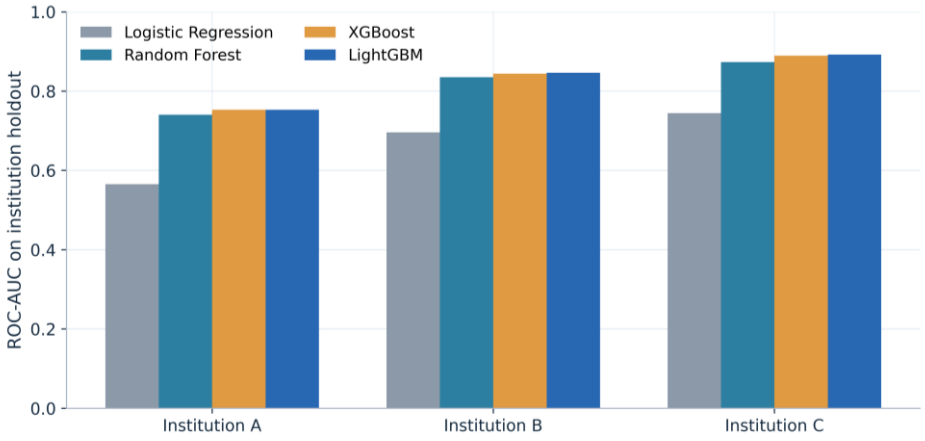


Figure 8. Locked-test and institution-level holdout discrimination across candidate models.

5.7. Explanation quality and bounded counterfactuals

Model-agnostic surrogate fidelity differed across 750 validation observations (Table 9): Logistic Regression had R2 = 0.997, Random Forest 0.887, LightGBM 0.823 and XGBoost 0.815. Thus, XGBoost, not LightGBM, had the lower of the two boosting-model surrogate-fidelity scores. LightGBM top-five permutation Jaccard was 0.667, versus 1.000 for the other models, indicating less agreement between the two ranking methods; its separate 300-case SHAP half-sample top-five Jaccard was 1.000. Expanded SHAP analyses used the same 300 validation observations for all four models, with maximum probability-reconstruction error below 0.000001 (Table 9A). Figures 9 and 10 present updated LightGBM and XGBoost attributions. SHAP magnitudes are not directly comparable across model-specific log-odds and probability scales.

Table 9. Model-agnostic explanation-quality measures

Model	Surrogate R^2	Top-five Jaccard	Sparsity (90%)	n
Logistic Regression	0.997	1.000	0.692	750
Random Forest	0.887	1.000	0.615	750
XGBoost	0.815	1.000	0.654	750
LightGBM	0.823	0.667	0.654	750

Table 9A. Expanded SHAP additivity and half-sample agreement (300 validation observations per model)

Model	SHAP n	Attribution scale	Maximum probability	Half-sample top-5

			error	Jaccard
Logistic Regression	300	log_odds	1.38e-08	0.6667
Random Forest	300	probability	7.71e-08	0.6667
XGBoost	300	log_odds	8.21e-07	0.6667
LightGBM	300	log_odds	7.11e-08	1

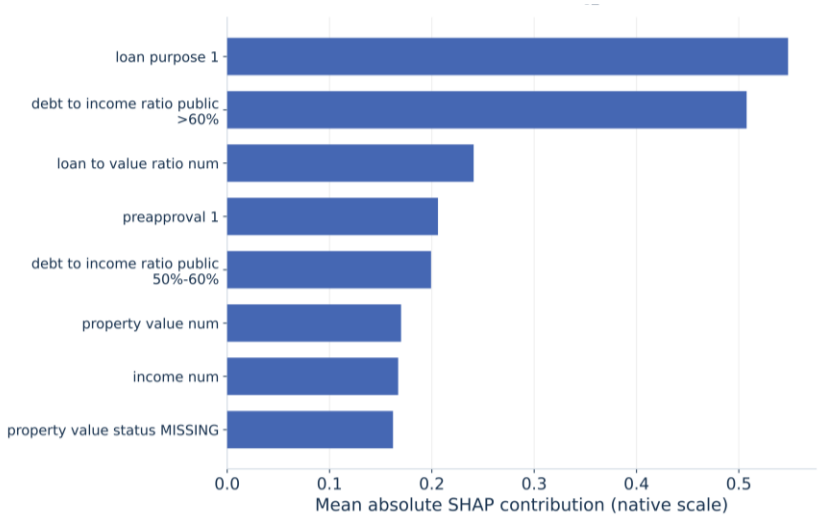


Figure 9. LightGBM SHAP attribution on 300 validation observations (log-odds output).

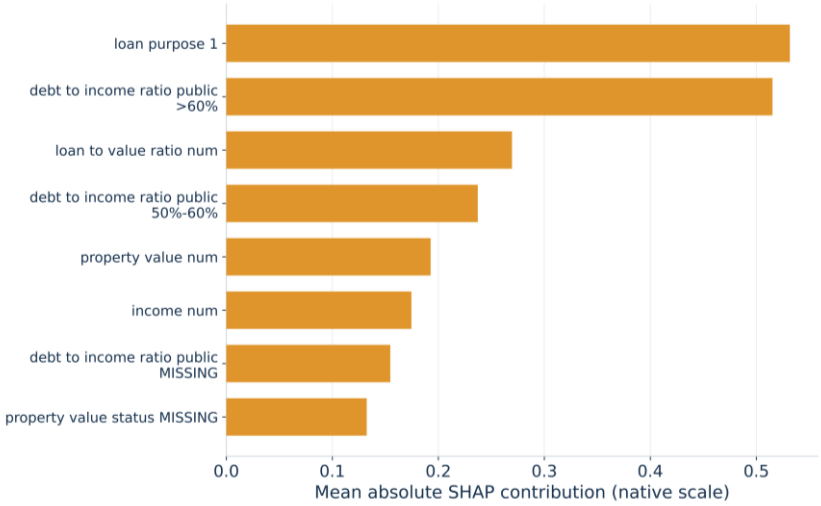


Figure 10. XGBoost native SHAP attribution on 300 validation observations (log-odds output).

Bounded counterfactual searches found reversals for 0/15 initially denied Logistic Regression cases, 2/19 Random Forest cases, 5/23 XGBoost cases and 3/21 LightGBM cases (Table 10). These correspond to 0%, 10.5%, 21.7% and 14.3% of initially denied cases, respectively. Counterfactual success was deliberately excluded from the original Explainability dimension, which measures surrogate fidelity, permutation agreement and attribution sparsity, because the bounded search did not establish valid, lawful or feasible borrower recourse. Consequently, Explainability scores describe a narrower technical construct and do not certify actionable adverse-action reasons.

Table 10. Bounded counterfactual reversal counts and success among initially denied cases

Model	Examined n	Initially denied n	Reversals found	Success per denied
Logistic Regression	60	15	0	0
Random Forest	60	19	2	0.1053
XGBoost	60	23	5	0.2174
LightGBM	60	21	3	0.1429

5.8. Fairness disparities and protected-group calibration

The recorded demographic-parity (DP) gaps were substantial, not small: 0.27636 for Logistic Regression, 0.26107 for Random Forest, 0.28683 for XGBoost and 0.28053 for LightGBM (Table 11), equivalent to approximately 26-29 percentage-point differences between reported group selection rates. At the original DP upper scoring anchor of 0.15, all models received a zero normalized DP component; this anchor was exploratory, not a regulatory standard. Increasing only the DP anchor to 0.30 raised the fairness scores from 5.61 to 8.24 (Logistic Regression), 16.17 to 20.50 (Random Forest), 10.36 to 11.82 (XGBoost) and 11.55 to 13.71 (LightGBM; Table 11B). With race groups restricted to n >= 1,000, original DP gaps were 0.27636, 0.19611, 0.21621 and 0.21451, respectively; matched-policy DP gaps were 0.20269, 0.21179, 0.21611 and 0.21451 (Table 11A). The groups and thresholds therefore affect comparisons. Figure 11 depicts group-rate support, whereas Figure 12 describes subgroup calibration; none of these descriptive gaps establishes legal discrimination.

Table 11. Original-threshold fairness gaps (all reported groups) and race DP sensitivity (group n >= 1,000)

Model	DP gap	EO gap	FPR gap	Restricted DP gap
Logistic Regression	0.2764	0.1248	0.2539	0.27636
Random Forest	0.2611	0.1039	0.1234	0.19611
XGBoost	0.2868	0.1058	0.1476	0.21621
LightGBM	0.2805	0.1014	0.1466	0.21451

Table 11A. Race-group disparities under original and matched thresholds (each included group n >= 1,000)

Model	Policy	Groups included	DP gap	EO gap	FPR gap
Logistic Regression	Original	8	0.2764	0.1248	0.2539
Logistic Regression	Matched	8	0.2027	0.1109	0.1696
Random Forest	Original	8	0.1961	0.1033	0.1234
Random Forest	Matched	8	0.2118	0.1064	0.1448
XGBoost	Original	8	0.2162	0.1058	0.1476
XGBoost	Matched	8	0.2161	0.1059	0.1471
LightGBM	Original	8	0.2145	0.1014	0.1466
LightGBM	Matched	8	0.2145	0.1014	0.1466

Table 11B. Research DP-anchor sensitivity (0.15 versus 0.30; EO/FPR anchors fixed)

Model	Anchor scenario	Fairness score /100	technical mean /100
Logistic Regression	0.15 DP anchor	5.6096	49.8161
Random Forest	0.15 DP anchor	16.1691	64.3620
XGBoost	0.15 DP anchor	10.3557	63.2088
LightGBM	0.15 DP anchor	11.5503	59.2543
Logistic Regression	0.30 DP anchor	8.2359	50.3414
Random Forest	0.30 DP anchor	20.4951	65.2272
XGBoost	0.30 DP anchor	11.8195	63.5016
LightGBM	0.30 DP anchor	13.7136	59.6870

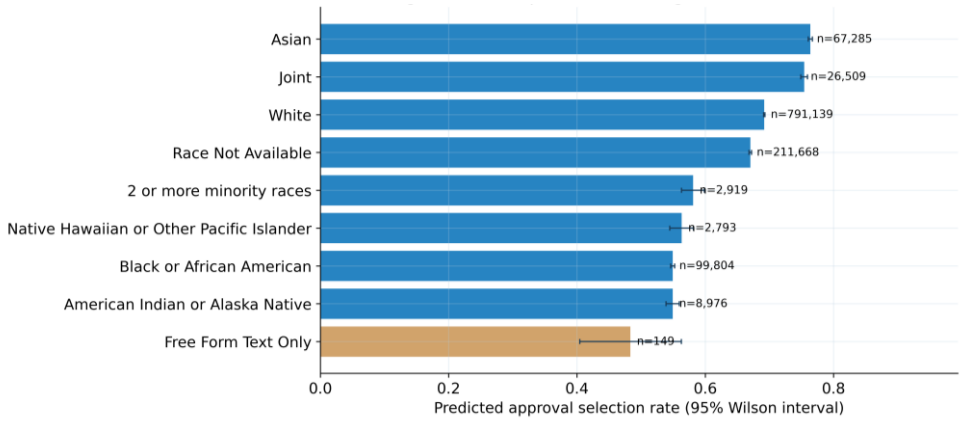


Figure 11. LightGBM predicted approval-selection rates and subgroup support intervals.

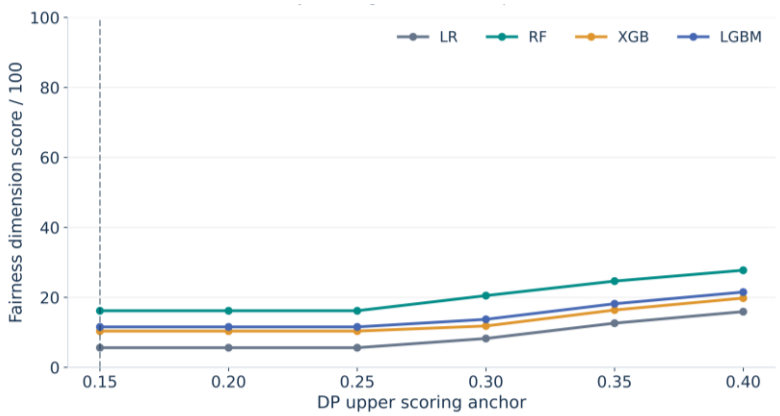


Figure 11A. Sensitivity of normalized fairness scores to the demographic-parity anchor (R6).

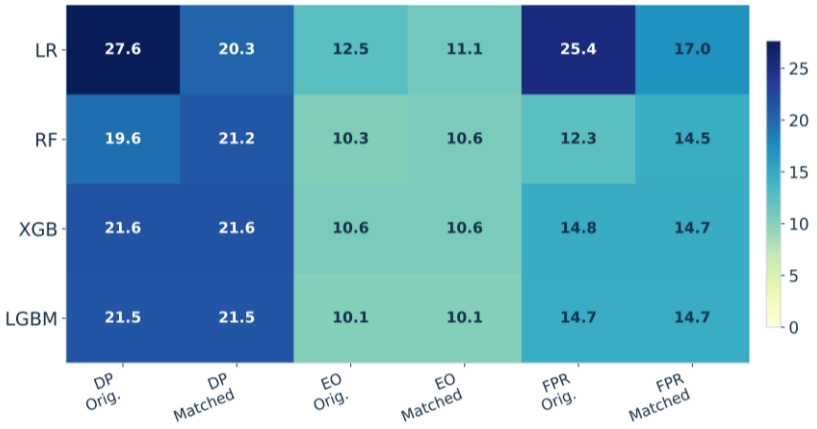


Figure 11B. Race-group disparity gaps at original and matched thresholds (n >= 1,000).

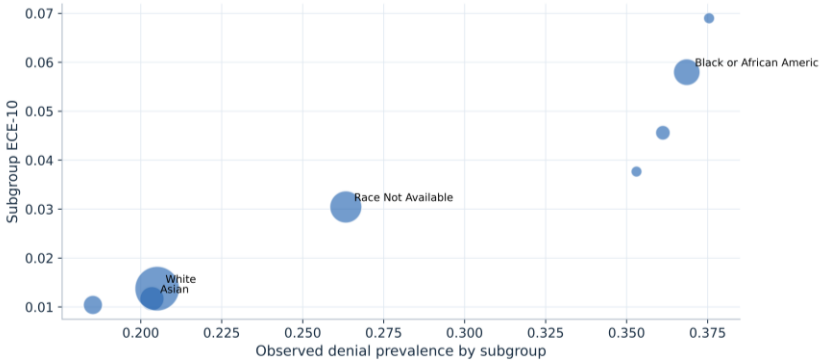


Figure 12. LightGBM subgroup predicted-denial calibration by protected-group category.

5.9. Training stability and perturbation robustness

Across 20 training refits, Random Forest had the lowest ROC-AUC coefficient of variation (0.000781), followed by XGBoost (0.001153), LightGBM (0.001237) and Logistic Regression (0.017725; Table 12). Stability differs from perturbation robustness: Logistic Regression had the highest anchored Robustness score (76.33), whereas Random Forest ranked highest for repeated-fit stability (98.70). In the separate 20,000-case validation-only input stress, LightGBM decision-flip rates were 13.995% under 10% missingness and 25.495% under 25% missingness (Table 13). The original 15% noise condition inadvertently used 5% noise; the replay used 15% and preserved the other four scenario results. Figures 13 and 14 display validation stress flip rates and nonnegative ROC-AUC losses; Table 13A reports a separate matched-threshold sensitivity sample. None of these figures is a locked-test stress evaluation.

Table 12. Training stability based on 20 repeated fits

Model	Refits	CV ROC-AUC	CV PR-AUC	CV Brier
LightGBM	20	0.00124	0.00197	0.00353
Logistic Regression	20	0.01772	0.04740	0.08096
Random Forest	20	0.00078	0.00130	0.00180
XGBoost	20	0.00115	0.00171	0.00305

An exploratory 100,000-training / 40,000-validation sample showed lower ROC-AUC after removing loan-to-value ratio, property value, income and debt-to-income ratio (Table 12A). This reduced-cost ablation demonstrates predictive dependence rather than independently proving decision-time leakage.

Table 12A. Exploratory ROC-AUC change after removing four underwriting-related predictors

Model	Train n	Validation n	Full ROC-AUC	Restricted ROC-AUC	Delta AUC
Logistic Regression	100,000	40,000	0.8246	0.7249	-0.0997
Random Forest	100,000	40,000	0.8608	0.7488	-0.1120
XGBoost	100,000	40,000	0.8628	0.7508	-0.1120
LightGBM	100,000	40,000	0.8654	0.7529	-0.1124

Table 13. Original-threshold validation-only missingness perturbation responses (n = 20,000)

Model	Stress	Nonnegative AUC loss	Decision-flip rate	Validation n
Logistic Regression	missing_10pct	0.0223	0.1283	20,000
Random Forest	missing_10pct	0.0513	0.1194	20,000
XGBoost	missing_10pct	0.0569	0.1241	20,000
LightGBM	missing_10pct	0.0756	0.1399	20,000
Logistic Regression	missing_25pct	0.0579	0.2344	20,000
Random Forest	missing_25pct	0.1256	0.2423	20,000
XGBoost	missing_25pct	0.1255	0.2344	20,000
LightGBM	missing_25pct	0.1565	0.2550	20,000

Table 13A. Sensitivity of validation-input stress flips to original versus matched thresholds

Model	Stress	Original flip rate	Matched flip rate	Matched - original
LightGBM	categorical_10pct	0.1145	0.1145	0
LightGBM	missing_10pct	0.1885	0.1885	0

LightGBM	missing_25pct	0.3245	0.3245	0
LightGBM	noise_15pct	0.1081	0.1081	0
LightGBM	noise_5pct	0.0915	0.0915	0
Logistic Regression	categorical_10pct	0.1197	0.1102	-0.0095
Logistic Regression	missing_10pct	0.1260	0.1172	-0.0087
Logistic Regression	missing_25pct	0.2338	0.2870	0.0532
Logistic Regression	noise_15pct	0.0091	0.0148	0.0057
Logistic Regression	noise_5pct	0.0031	0.0056	0.0024
Random Forest	categorical_10pct	0.1567	0.1591	0.0024
Random Forest	missing_10pct	0.2041	0.2001	-0.0040
Random Forest	missing_25pct	0.3762	0.3532	-0.0231
Random Forest	noise_15pct	0.0382	0.0410	0.0028
Random Forest	noise_5pct	0.0192	0.0210	0.0018
XGBoost	categorical_10pct	0.0932	0.0933	5.00e-05
XGBoost	missing_10pct	0.1796	0.1794	-0.0002
XGBoost	missing_25pct	0.3219	0.3220	0.0001
XGBoost	noise_15pct	0.1197	0.1194	-0.0003
XGBoost	noise_5pct	0.1158	0.1155	-0.0002

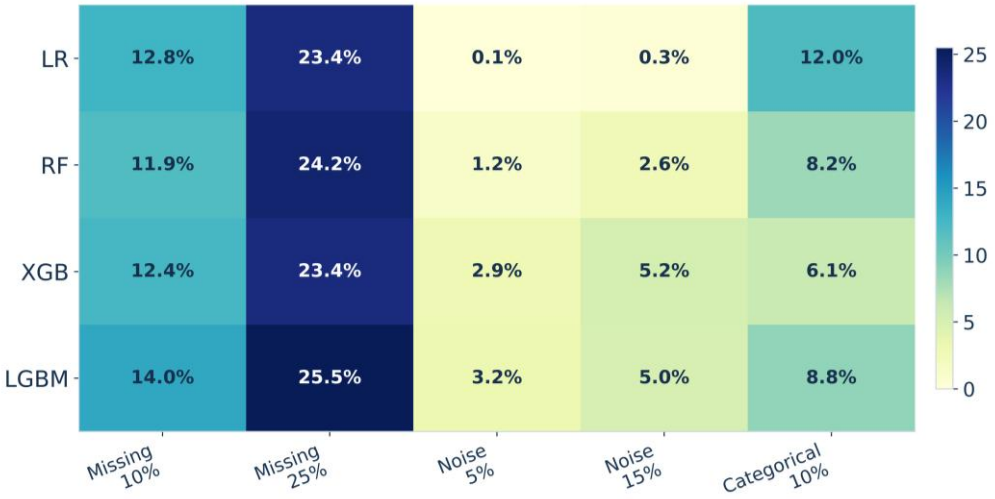


Figure 13.Validation-only perturbation decision-flip rates at original thresholds.

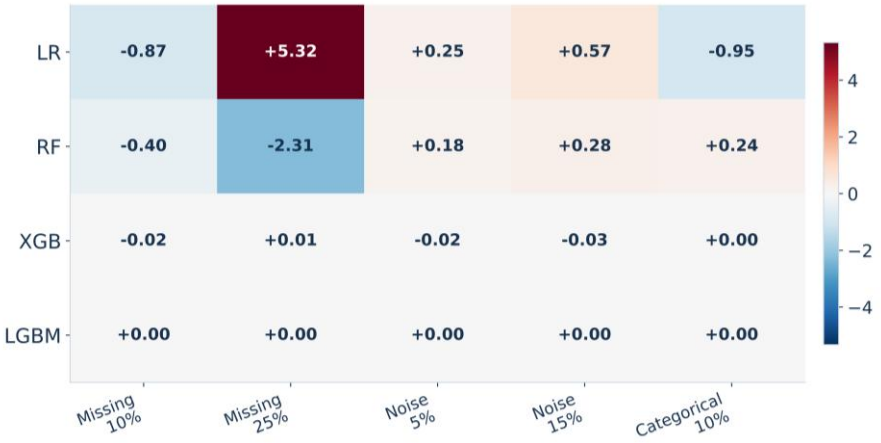


Figure 13A. Difference in stress-induced flip rates after matching validation-defined thresholds.

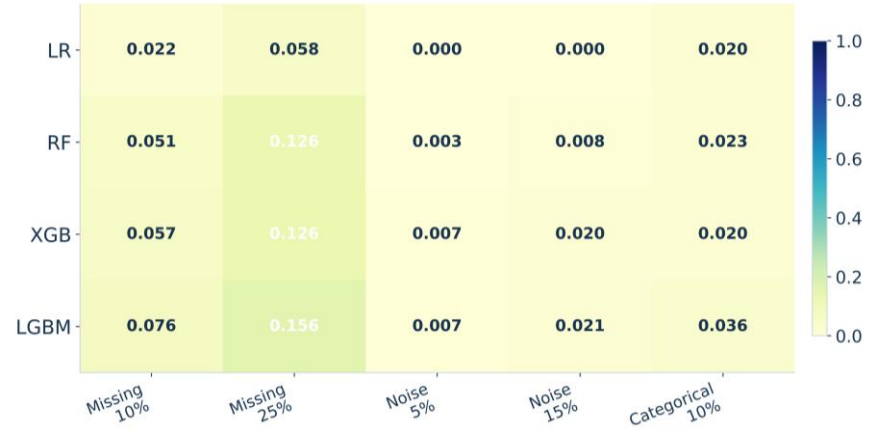


Figure 14.Nonnegative ROC-AUC loss under validation-only input perturbations.

5.10. Measurable BAIMRS technical profiles

Governance Readiness remained unscored, while five technical dimensions were measured for four core models. After correcting the 15% numeric-noise stress, the provisional five-dimension means were Random Forest 64.362, XGBoost 63.209, LightGBM 59.254 and Logistic Regression 49.816 (Table 14). Random Forest was the nominal research-anchor leader, but XGBoost ranked first at the 0.75 anchor-tolerance scale (Table 14B); therefore, the leading two candidates should be described as closely ranked and sensitive to normalization policy, not as having an unconditional winner. Figure 15 shows the technical profiles. The fairness scores remain low partly because all DP gaps exceed the research 0.15 normalization ceiling. The required independent governance and auditability evidence was not available for a formal deployment composite.

Table 14. Provisional five-dimensional BAIMRS technical profiles (0-100)

Model	Predictive Performance	Explainability	Fairness	Stability	Robustness	Provisional mean
Logistic Regression	26.1753	89.6599	5.6096	51.3059	76.3299	49.8161
Random Forest	62.3859	83.4247	16.1691	98.7042	61.1261	64.3620
XGBoost	67.2623	82.3065	10.3557	98.0300	58.0895	63.2088
LightGBM	67.3025	71.4630	11.5503	97.7567	48.1990	59.2543

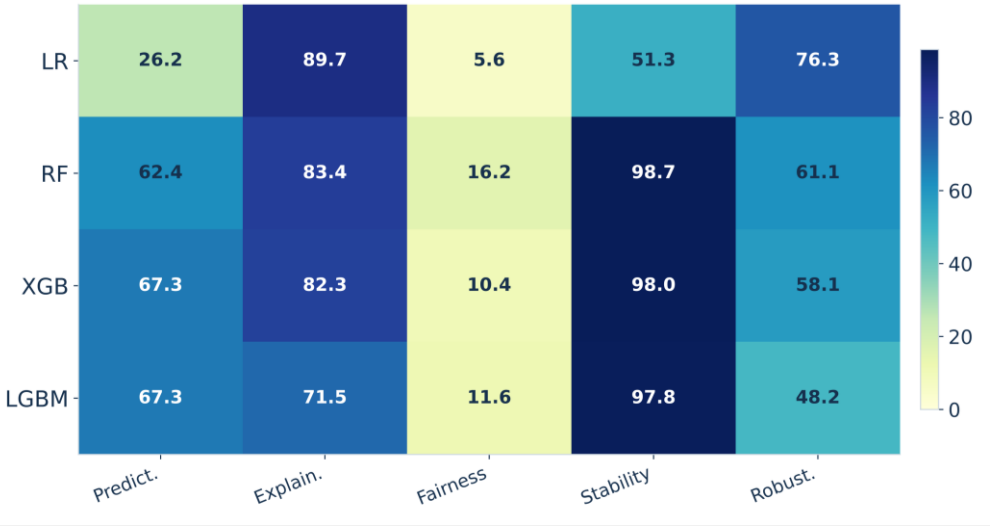


Figure 15. Five-dimensional BAIMRS technical profiles; governance is not scored.

LightGBM led locked-test ROC-AUC, whereas Random Forest had the highest nominal provisional technical mean. This reversal is descriptive, depends on research scoring anchors, and does not constitute a deployment recommendation: the required governance and auditability gates remain unverified.

5.11. Weight sensitivity, anchor tolerance and ablation

In 3,000 reproducible Dirichlet-weight draws recomputed using Robustness evidence, the provisional rank-one frequencies were Random Forest 73.00%, XGBoost 20.33%, Logistic Regression 6.53% and LightGBM 0.13% (Table 14C; Figure 16). These are descriptive fractions over exploratory weights, not the probability of regulatory acceptance or banking deployment. At the stricter anchor-tolerance scale of 0.75, XGBoost achieved a provisional mean of 64.443 compared with Random Forest 63.907; nominal-scale results instead favored Random Forest 64.362 over XGBoost 63.209 (Table 14B). Conditional technical-score p2.5-p97.5 bands from 20 limited refits were 64.335-64.379 for Random Forest and 63.176-63.239 for XGBoost (Table 14A); these are not full 95% confidence intervals. Leave-one-dimension-out rankings are shown in Figure 17.

Table 14A. Conditional variation of technical means across 20 training refits (not 95% CIs)

Model	Conditional p2.5	Conditional p97.5	SD
Logistic Regression	48.3331	50.0063	0.5099
Random Forest	64.3355	64.3790	0.0124
XGBoost	63.1765	63.2386	0.0198
LightGBM	59.2152	59.2965	0.0229

Table 14B. Technical means under alternative research-anchor tolerances

Anchor scale	Model	Technical mean	Rank
0.7500	Logistic Regression	47.0217	4
0.7500	Random Forest	63.9068	2
0.7500	XGBoost	64.4431	1
0.7500	LightGBM	61.2801	3
1	Logistic Regression	49.8161	4
1	Random Forest	64.3620	1
1	XGBoost	63.2088	2
1	LightGBM	59.2543	3
1.2500	Logistic Regression	49.9862	4
1.2500	Random Forest	62.9563	1
1.2500	XGBoost	62.0337	2
1.2500	LightGBM	58.8701	3

Table 14C. Provisional rank-one fractions across 3,000 research-weight draws

Model	Exploratory rank-one frequency	Weight draws
Logistic Regression	0.0653	3,000
Random Forest	0.7300	3,000
XGBoost	0.2033	3,000
LightGBM	0.0013	3,000

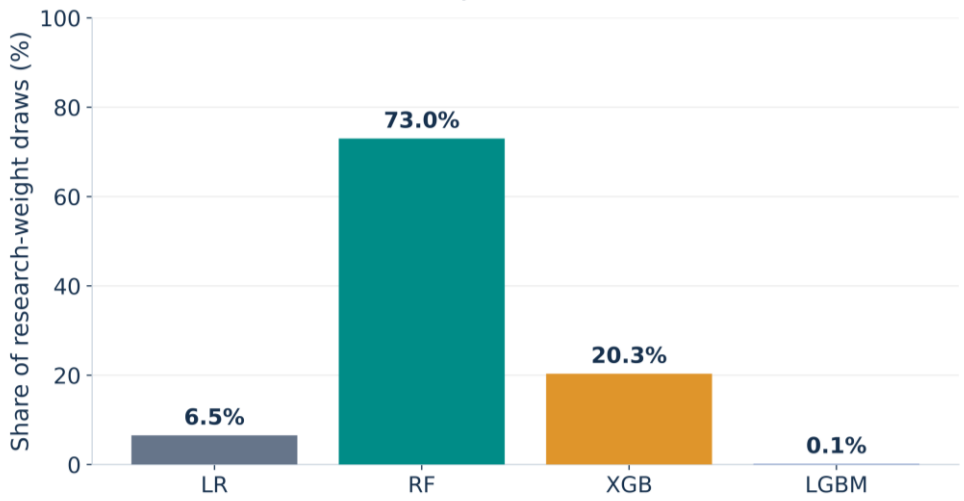


Figure 16. Conditional rank-one frequency under 3,000 research-weight draws.

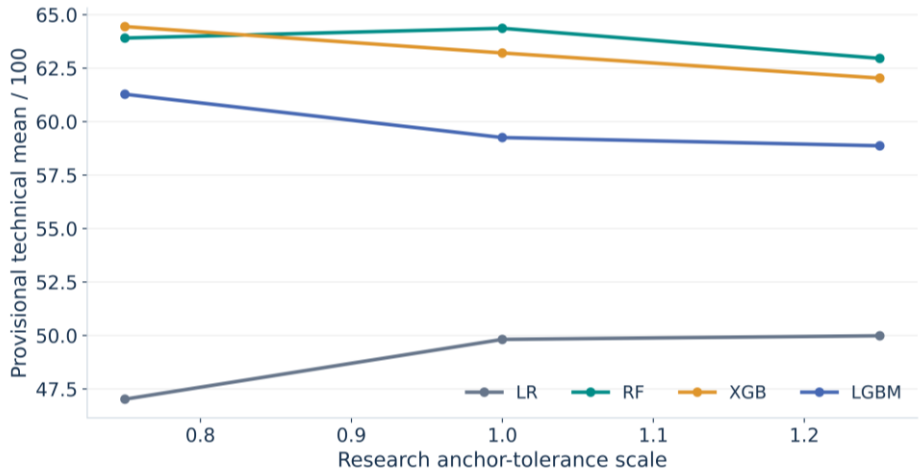


Figure 16A. Ranking sensitivity to the research-anchor tolerance after robustness correction.

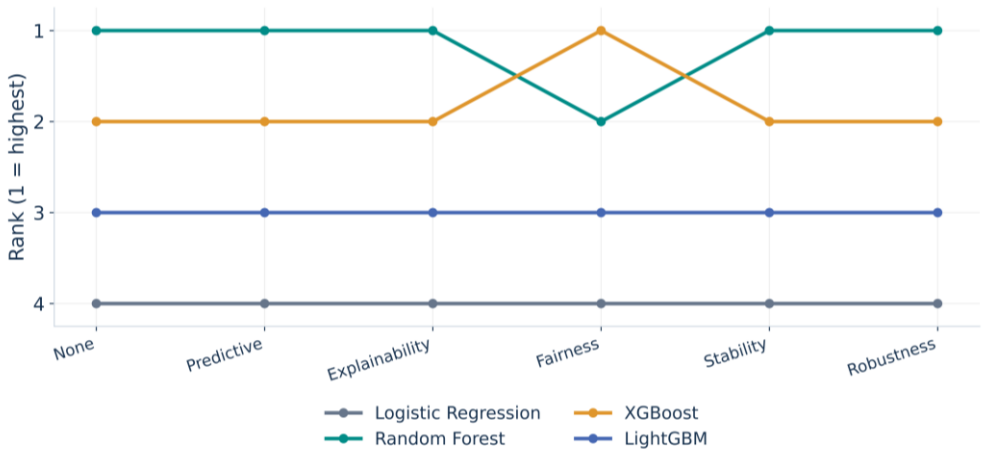


Figure 17. Leave-one-technical-dimension-out rank sensitivity.

The weight analysis is a conditional ranking over the weight vectors of the explorations, and has nothing to do with the probability that a banking committee or a banking regulator would recommend a candidate. The change at the 0.75 anchor point also implies the nature of the descriptive leader will vary because it is determined by statements of scoring tolerances that are not elicited separately.

5.12. Governance evidence, auditability and eligibility

Source, model, preprocessing and prediction checksums were recorded, but independent replay attestation, approved anchor/weight policies and independent governance evaluations were unavailable. Accordingly, no formal BAIMRS-Core score or verified deployment eligibility could be assigned to any of the four models (Table 15; Figure 18). The technical means are descriptive research summaries, not banking approval or regulatory certification.

Table 15.Technical means and unavailable governance

Model	Technical mean
Logistic Regression	49.8161
Random Forest	64.3620
XGBoost	63.2088
LightGBM	59.2543

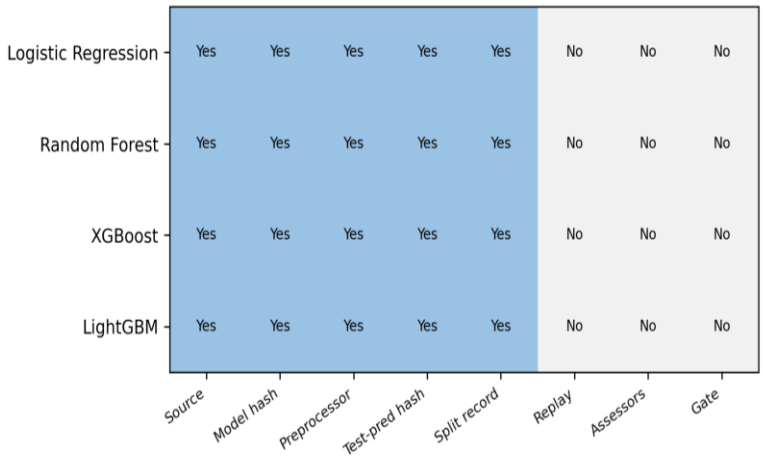


Figure 18. Current evidence-gate status: verified artifact traces versus missing replay and assessor validation.

6. Discussion

The analysis shows that predictive AUC leadership and multi-dimensional technical assessments answer different questions. LightGBM had the highest numerical AUC, but its paired difference from XGBoost was not statistically distinguishable from zero. Random Forest had the highest nominal equal-weight technical mean (64.362), closely followed by XGBoost (63.209), whereas XGBoost led under the 0.75 anchor-tolerance scale. Logistic Regression had the highest anchored Robustness dimension score (76.33), while Random Forest was strongest in repeated-fit Stability (98.70); these are different constructs. Because scoring tolerances were research-selected and governance evidence was missing, the study does not establish a single deployment-suitable winner.

Trade-offs between uncertainty and loss preferences aspects of multiobjective credit-scoring work have been highlighted [65]. Profit-sensitive modelling also suggests that traditional predictive measures are not necessarily a true reflection of decision consequences [66] and profit-based score evaluation explicitly alters the optimization objective [67]. Unlike cost-based approaches [66], [67], the present BAIMRS experiments measure technical model-risk dimensions only; they do not quantify organizational welfare, profitability, loan losses or economic return. Real lending environments are supported by evidence related to lender automation and racial disparities [68] and across-dataset fairness audits [69] stress the need to not give an “accuracy only” reading of the problem. The work on bias-mitigation research [70] provides a potential pathway to the further development of the existing limited alternative-model screen. Issues on transparency and dynamic learning are highlighted in reviews of AI credit assessment [71] as well. Validating the model is linked to governance

documentation via responsible financial AI scholarship [72] and there are additional options for traceability beyond the current static pipeline that can be shown with secure distributed credit models [73]. Explanation-guided learning emphasizes the importance of explanations in not just diagnostics, but also in the learning process of improving models [74]. Last but not least, tests of performance-decomposition for credit scoring [75] indicate that differences in the success of the predictions should be interpreted with caution as causal and operational, and not generalized.

The comparison in Sections 5.10 and 5.11 does not claim that the AUC-leading model is an operationally superior choice; it documents how provisional rankings depend on selected technical weights and normalization anchors. The tests do not isolate model complexity as a causal factor, identify the lender's original decision rules or estimate welfare effects of alternative lending policies.

The HMDA label measures historical lender action, not borrower repayment, default or whether each denial was normatively appropriate. Public data omit applicant credit scores and other underwriting information, and prospective availability of several predictor fields was not independently confirmed. Measured group gaps are therefore descriptive screening signals rather than evidence of unlawful disparate treatment. For example, a race category with fewer than 1,000 records can determine an all-category DP gap while being excluded from the n >= 1,000 sensitivity analysis. Validation-defined matched thresholds improve policy comparability but do not remove label bias or missing confounders. Neither SHAP attribution nor bounded numerical counterfactuals establish legally adequate adverse-action reasons.

A practical banking assessment should maintain the traceability of the predictions, a separate validation strategy, clear definitions of metrics, a priori tolerances and an independent examination of the results of calibration and calibration subgroups. BAIMRS will be tailored to ensure that these requirements are traceable and that missing evidence is considered to be a blocking condition and not imputing favourable scores. As deliberately non-compensatory, you can't simply compensate for missing auditability, policy authorisation or failure on the fairness criterion with good AUC. The gate as implemented, does not declare that the candidate models do not comply with a statute, nor certify that other systems comply with law.

For operational use, model inventories, individual protected-variable and proxy coding would need to be approved and versioned, individual coding would need to be validated and documented, there would need to be an approved, versioned reason-code review by qualified personnel, and there would need to be institution-specific tolerances for the model and subgroup outcomes. The proposed control does not result in a formal eligibility pass because it is only evidence that can be used for reporting to the Provenance Check but is not necessary for an end-to-end decision replay.

The reference anchors and weights were researcher-selected, not elicited through an approved institutional process. Independent replay attestation and governance-assessor ratings were unavailable. Twenty sampled training refits provide limited variability information rather than full-score uncertainty; the reported technical-score percentile ranges are conditional sensitivity bands, not formal confidence intervals. Expanded SHAP verification involved 300 shared validation observations per model, rather than the original 80, but explanations were not independently validated as adverse-action reasons. Bounded counterfactual reversals also lacked verified legal or practical actionability. No cross-year HMDA test, borrower-default outcome, independent assessor agreement or validated principal denial reasons were available.

Future research should examine cohorts over the HMDA years, test the sensitivity of the framework to creditworthiness features and create a consensus on reason-code concordance on real mortgage-performance labels. The governance rubric and scoring anchors as well as policy thresholds could be validated by independent experts (with ethics requirements applicable). If a formal deployment suitability claim is to be made, then it must be independently verified (replayed) and independently verified (logged) at the replay/decision level. The BAIMRS-G pathway then evaluates evidence-constrained language-model explanations, factuality, prompt sensitivity and human oversight, but not here, does there exist any comparable reliability results for GenAI.

If a bank began using this assessment, then they would need to preserve a model and a preprocessed version of it, the threshold, and the evidence record and predecided escalation criteria for each production decision. Ongoing monitoring could involve comparing probability calibration, support of sub groups, institution shifts and input drift/explanation consistency with validated tolerances, and independent challenge after a material change. These recommendations are operational requirements which have been developed based on the framework design and are not validated during the current assessment.

7. Conclusion

BAIMRS was developed to assess historical mortgage-denial prediction across discrimination, explainability,

fairness, stability, robustness and governance readiness without compensating for missing audit evidence. The retrospective 2025 HMDA cohort included 9,360,942 application decisions, four primary models, a monotone LightGBM benchmark, one locked test set and three same-year institution holdouts. LightGBM achieved the highest locked-test ROC-AUC (0.8729), but its 0.000179 advantage over XGBoost had a paired interval crossing zero. At Institution A its AUC decreased to 0.7530, and at Institution C its calibration ECE10 rose to 0.1129, underscoring transfer limitations. After correcting the numeric-noise stress protocol, Random Forest had the highest nominal provisional technical mean (64.362), followed by XGBoost (63.209); XGBoost led at anchor scale 0.75, so neither is an unconditional winner. The original demographic-parity gaps reached approximately 26-29 percentage points, and expanded SHAP checks used 300 validation observations per model. Governance and independent replay requirements remained unmet, precluding formal deployment ranking. These results support transparent comparative model-risk diagnostics, not lender-default prediction, economic-return assessment, actionable recourse or regulatory certification. Future work should validate decision-time features, temporal transfer, approved policy thresholds, adverse-action reasons, governance evaluation and borrower-performance outcomes.

References

1. H. Ayari, P. R. Guetari, and P. N. Kraïem, “Machine learning powered financial credit scoring: A systematic literature review,” Artificial Intelligence Review, vol. 59, Art. no. 13, 2026. doi: 10.1007/s10462-025-11416-2.
2. R. Bahlool, N. Hewahi, and W. Elmedany, “Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review,” Journal of Risk and Financial Management, vol. 19, no. 2, Art. no. 104, 2026. doi: 10.3390/jrfm19020104.
3. A. A. Montevechi, R. de Carvalho Miranda, A. L. Medeiros, and J. A. B. Montevechi, “Advancing credit risk modelling with machine learning: A comprehensive review of the state-of-the-art,” Engineering Applications of Artificial Intelligence, vol. 137, Art. no. 109082, 2024. doi: 10.1016/j.engappai.2024.109082.
4. L. Gambacorta, Y. Huang, H. Qiu, and J. Wang, “How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm,” Journal of Financial Stability, vol. 73, Art. no. 101284, 2024. doi: 10.1016/j.jfs.2024.101284.
5. P. Giudici, M. Centurelli, and S. Turchetta, “Artificial intelligence risk measurement,” Expert Systems with Applications, vol. 235, Art. no. 121220, 2024. doi: 10.1016/j.eswa.2023.121220.
6. A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, “Predictably unequal? The effects of machine learning on credit markets,” The Journal of Finance, vol. 77, no. 1, pp. 5–47, 2022. doi: 10.1111/jofi.13090.
7. R. Bartlett, A. Morse, R. Stanton, and N. Wallace, “Consumer-lending discrimination in the fintech era,” Journal of Financial Economics, vol. 143, no. 1, pp. 30–56, 2022. doi: 10.1016/j.jfineco.2021.05.047.
8. N. Bhutta, A. Hizmo, and D. Ringo, “How much does racial bias affect mortgage lending? Evidence from human and algorithmic credit decisions,” The Journal of Finance, vol. 80, no. 3, pp. 1463–1496, 2025. doi: 10.1111/jofi.13444.
9. S. Das, R. Stanton, and N. Wallace, “Algorithmic fairness,” Annual Review of Financial Economics, vol. 15, pp. 565–593, 2023. doi: 10.1146/annurev-financial-110921-125930.
10. M. von Zahn, J. Zacharias, M. Lowin, J. Chen, and O. Hinz, “Navigating AI conformity: A design framework to assess fairness, explainability, and performance,” Electronic Markets, vol. 35, Art. no. 24, 2025. doi: 10.1007/s12525-025-00770-2.
11. S. Cosma, G. Rimo, and G. Torluccio, “Knowledge mapping of model risk in banking,” International Review of Financial Analysis, vol. 89, Art. no. 102800, 2023. doi: 10.1016/j.irfa.2023.102800.
12. M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek, “Transparency, auditability, and explainability of machine learning models in credit scoring,” Journal of the Operational Research Society, vol. 73, no. 1, pp. 70–90, 2022. doi: 10.1080/01605682.2021.1922098.
13. V. Moscato, A. Picariello, and G. Sperlí, “A benchmark of machine learning approaches for credit score prediction,” Expert Systems with Applications, vol. 165, Art. no. 113986, 2021. doi: 10.1016/j.eswa.2020.113986.
14. E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, “Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects,” European Journal of Operational Research, vol. 297, no. 3, pp. 1178–1192, 2022. doi: 10.1016/j.ejor.2021.06.053.
15. N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, “Explainable machine learning in credit risk management,” Computational Economics, vol. 57, no. 1, pp. 203–216, 2021. doi: 10.1007/s10614-020-10042-0.
16. Y. Chen, R. Calabrese, and B. Martin-Barragan, “Interpretable machine learning for imbalanced credit scoring datasets,” European Journal of Operational Research, vol. 312, no. 1, pp. 357–372, 2024. doi: 10.1016/j.ejor.2023.06.036.
17. V. Chang, Q. A. Xu, S. H. Akinloye, V. Benson, and K. Hall, “Prediction of bank credit worthiness through credit

risk analysis: An explainable machine learning study," Annals of Operations Research, vol. 354, pp. 247–271, 2025. doi: 10.1007/s10479-024-06134-x.
18. H.-W. Teng, M.-H. Kang, I.-H. Lee, and L.-C. Bai, "Bridging accuracy and interpretability: A rescaled cluster-then-predict approach for enhanced credit scoring," International Review of Financial Analysis, vol. 91, Art. no. 103005, 2024. doi: 10.1016/j.irfa.2023.103005.
19. Y. Liu, F. Huang, L. Ma, Q. Zeng, and J. Shi, "Credit scoring prediction leveraging interpretable ensemble learning," Journal of Forecasting, vol. 43, no. 2, pp. 286–308, 2024. doi: 10.1002/for.3033.
20. K. Idbenjra, K. Coussement, and A. De Caigny, "Investigating the beneficial impact of segmentation-based modelling for credit scoring," Decision Support Systems, vol. 179, Art. no. 114170, 2024. doi: 10.1016/j.dss.2024.114170.
21. L. Wang, Z. Yu, J. Ma, X. Chen, and C. Wu, "A two-stage interpretable model to explain classifier in credit risk prediction," Journal of Forecasting, vol. 44, no. 7, pp. 2132–2150, 2025. doi: 10.1002/for.3288.
22. A. Khan and S. K. Ghosh, "Machine assistance for credit approval? Random wheel can recommend and explain," Expert Systems with Applications, vol. 215, Art. no. 119231, 2023. doi: 10.1016/j.eswa.2022.119231.
23. D. Chen, J. Ye, and W. Ye, "Interpretable selective learning in credit risk," Research in International Business and Finance, vol. 65, Art. no. 101940, 2023. doi: 10.1016/j.ribaf.2023.101940.
24. J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: A systematic literature review," Artificial Intelligence Review, vol. 57, Art. no. 216, 2024. doi: 10.1007/s10462-024-10854-8.
25. A. Gramegna and P. Giudici, "SHAP and LIME: An evaluation of discriminative power in credit risk," Frontiers in Artificial Intelligence, vol. 4, Art. no. 752558, 2021. doi: 10.3389/frai.2021.752558.
26. P. E. De Lange, B. Melsom, C. B. Vennerød, and S. Westgaard, "Explainable AI for credit assessment in banks," Journal of Risk and Financial Management, vol. 15, no. 12, Art. no. 556, 2022. doi: 10.3390/jrfm15120556.
27. C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations," Decision Support Systems, vol. 152, Art. no. 113647, 2022. doi: 10.1016/j.dss.2021.113647.
28. G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," Information Fusion, vol. 76, pp. 89–106, 2021. doi: 10.1016/j.inffus.2021.05.009.
29. A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: Contrastive explanations and consequential recommendations," ACM Computing Surveys, vol. 55, no. 5, Art. no. 95, 2023. doi: 10.1145/3527848.
30. M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," ACM Computing Surveys, vol. 55, no. 13s, Art. no. 295, 2023. doi: 10.1145/3583558.
31. R. Guidotti, "Counterfactual explanations and how to find them: Literature review and benchmarking," Data Mining and Knowledge Discovery, vol. 38, pp. 2770–2824, 2024. doi: 10.1007/s10618-022-00831-6.
32. R. M. B. de Oliveira, K. Sörensen, and D. Martens, "A model-agnostic and data-independent tabu search algorithm to generate counterfactuals for tabular, image, and text data," European Journal of Operational Research, vol. 317, no. 2, pp. 286–302, 2024. doi: 10.1016/j.ejor.2023.08.031.
33. E. Sepúlveda, F. Vandervorst, B. Baesens, and T. Verdonck, "Enhancing explainability in real-world scenarios: Towards a robust stability measure for local interpretability," Expert Systems with Applications, vol. 274, Art. no. 126922, 2025. doi: 10.1016/j.eswa.2025.126922.
34. F. S. Khan, S. S. Mazhar, K. Mazhar, D. A. AlSaleh, and A. Mazhar, "Model-agnostic explainable artificial intelligence methods in finance: A systematic review, recent developments, limitations, challenges and future directions," Artificial Intelligence Review, vol. 58, Art. no. 232, 2025. doi: 10.1007/s10462-025-11215-9.
35. W. J. Yeo, W. Van Der Heever, R. Mao, E. Cambria, R. Satapathy, and G. Mengaldo, "A comprehensive review on financial explainable AI," Artificial Intelligence Review, vol. 58, Art. no. 189, 2025. doi: 10.1007/s10462-024-11077-7.
36. S. Genovesi, J. M. Mönig, A. Schmitz, M. Poretschkin, M. Akila, M. Kahdan, et al., "Standardizing fairness-evaluation procedures: Interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans," AI and Ethics, vol. 4, pp. 537–553, 2024. doi: 10.1007/s43681-023-00291-8.
37. C. N. Nwafor, O. Nwafor, and S. Brahma, "Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach," Scientific Reports, vol. 14, Art. no. 25174, 2024. doi: 10.1038/s41598-024-75026-8.
38. N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," European Journal of Operational Research, vol. 297, no. 3, pp. 1083–1094, 2022. doi: 10.1016/j.ejor.2021.06.023.
39. Y. Chen, P. Giudici, K. Liu, and E. Raffinetti, "Measuring fairness in credit ratings," Expert Systems with Applications, vol. 258, Art. no. 125184, 2024. doi: 10.1016/j.eswa.2024.125184.
40. S. Agarwal, C. B. Muckley, and P. Neelakantan, "Countering racial discrimination in algorithmic lending: A case

for model-agnostic interpretation methods," Economics Letters, vol. 226, Art. no. 111117, 2023. doi: 10.1016/j.econlet.2023.111117.
41. A. C. B. Garcia, M. G. P. Garcia, and R. Rigobon, "Algorithmic discrimination in the credit domain: What do we know about it?," AI & Society, vol. 39, pp. 2059–2098, 2024. doi: 10.1007/s00146-023-01676-3.
42. Z. Liu and H. Liang, "Are credit scores gender-neutral? Evidence of mis-calibration from alternative and traditional borrowing data," Journal of Behavioral and Experimental Finance, vol. 47, Art. no. 101081, 2025. doi: 10.1016/j.jbef.2025.101081.
43. H. Giang Thi Thu, T. V. Doan, H.-B. Ban, and T. L. Quy, "An experimental study on fairness-aware machine learning for credit scoring problems," Digital Finance, vol. 8, Art. no. 35, 2026. doi: 10.1007/s42521-026-00202-6.
44. C. Mariscal, Y. Yustiawan, F. C. Rochim, and E. Tanuar, "Implementing and analyzing fairness in banking credit scoring," Procedia Computer Science, vol. 234, pp. 1492–1499, 2024. doi: 10.1016/j.procs.2024.03.150.
45. Z. Chen, J. M. Zhang, M. Hort, M. Harman, and F. Sarro, "Fairness testing: A comprehensive survey and analysis of trends," ACM Transactions on Software Engineering and Methodology, vol. 33, no. 5, Art. no. 137, 2024. doi: 10.1145/3652155.
46. T. D. Jui and P. Rivas, "Fairness issues, current approaches, and challenges in machine learning models," International Journal of Machine Learning and Cybernetics, vol. 15, pp. 3095–3125, 2024. doi: 10.1007/s13042-023-02083-2.
47. D. Pessach and E. Shmueli, "A review on fairness in machine learning," ACM Computing Surveys, vol. 55, no. 3, Art. no. 51, 2022. doi: 10.1145/3494672.
48. M. Ballegeer, M. Bogaert, and D. F. Benoit, "Evaluating the stability of model explanations in instance-dependent cost-sensitive credit scoring," European Journal of Operational Research, vol. 326, no. 3, pp. 630–640, 2025. doi: 10.1016/j.ejor.2025.05.039.
49. A. C. Bueff, M. Cytryński, R. Calabrese, M. Jones, J. Roberts, J. Moore, and I. Brown, "Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals," Expert Systems with Applications, vol. 202, Art. no. 117271, 2022. doi: 10.1016/j.eswa.2022.117271.
50. D. Nikolaidis and M. Doumpos, "Credit scoring with drift adaptation using local regions of competence," Operations Research Forum, vol. 3, Art. no. 67, 2022. doi: 10.1007/s43069-022-00177-1.
51. N. Kemmerzell, A. Schreiner, H. Khalid, M. Schalk, and L. Bordoli, "Towards a better understanding of evaluating trustworthiness in AI systems," ACM Computing Surveys, vol. 57, no. 9, Art. no. 218, 2025. doi: 10.1145/3721976.
52. D. Kowald, S. Scher, V. Pammer-Schindler, P. Müllner, K. Waxnegger, et al., "Establishing and evaluating trustworthy AI: Overview and research challenges," Frontiers in Big Data, vol. 7, Art. no. 1467222, 2024. doi: 10.3389/fdata.2024.1467222.
53. M. P. Mathen and A. Paul, "Toward an evolving framework for responsible AI for credit scoring in the banking industry," Journal of Information, Communication and Ethics in Society, vol. 23, no. 1, pp. 148–163, 2025. doi: 10.1108/JICES-08-2024-0122.
54. M. Golec and M. Alabduljalil, "Interpretable LLMs for credit risk: A systematic review and taxonomy," Expert Systems with Applications, vol. 306, Art. no. 130941, 2026. doi: 10.1016/j.eswa.2025.130941.
55. Consumer Financial Protection Bureau, "Equal Credit Opportunity Act (Regulation B)," 12 CFR pt. 1002, particularly §§ 1002.4, 1002.6 and 1002.9. [Online]. Available: https://www.consumerfinance.gov/rules-policy/regulations/1002/.
56. United States Code, "Discrimination in residential real estate-related transactions," 42 U.S.C. § 3605. [Online]. Available: https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title42-section3605.
57. Consumer Financial Protection Bureau, "Home Mortgage Disclosure (Regulation C)," 12 CFR pt. 1003. [Online]. Available: https://www.consumerfinance.gov/rules-policy/regulations/1003/.
58. Board of Governors of the Federal Reserve System, "Revised Guidance on Model Risk Management," Supervisory Letter SR 26-2, Apr. 17, 2026. [Online]. Available: https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm.
59. Federal Financial Institutions Examination Council, "2025 Snapshot National Loan-Level Dataset: Loan/Application Records (LAR)," HMDA Data Publication, 2026. [Online]. Available: https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2025.
60. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
61. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, M. Ma, et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, 2017. [Online]. Available:
62. https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html
63. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural

Information Processing Systems, vol. 30, 2017. [Online]. Available:
64. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
65. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
66. M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems, vol. 29, 2016. [Online]. Available:
67. https://papers.neurips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html
68. Y. Xu, G. Kou, Y. Peng, K. Ding, D. Ergu, and F. S. Alotaibi, "Profit- and risk-driven credit scoring under parameter uncertainty: A multiobjective approach," Omega, vol. 125, Art. no. 103004, 2024. doi: 10.1016/j.omega.2023.103004.
69. M.-J. Ariza-Garzón, J. Arroyo, M.-J. Segovia-Vargas, and A. Caparrini, "Profit-sensitive machine learning classification with explanations in credit risk: The case of small businesses in peer-to-peer lending," Electronic Commerce Research and Applications, vol. 67, Art. no. 101428, 2024. doi: 10.1016/j.elerap.2024.101428.
70. H. Mohammadnejad-Daryani, A. A. Taleizadeh, and D. Pamucar, "A novel profit-driven framework for model evaluation in credit scoring," Engineering Applications of Artificial Intelligence, vol. 137, Art. no. 109137, 2024. doi: 10.1016/j.engappai.2024.109137.
71. S. T. Howell, T. Kuchler, D. Snitkof, J. Stroebel, and J. Wong, "Lender automation and racial disparities in credit access," The Journal of Finance, vol. 79, no. 2, pp. 1457–1512, 2024. doi: 10.1111/jofi.13303.
72. C. Ellis, "The fairness illusion? A cross-dataset audit of accuracy and demographic bias in credit scoring based on machine learning," Journal of Risk and Financial Management, vol. 19, no. 9, Art. no. 674, 2026. doi: 10.3390/jrfm19090674.
73. J. Lin and L. Yang, "Mitigating algorithmic bias in credit scoring support systems through adversarial learning," Decision Support Systems, vol. 205, Art. no. 114653, 2026. doi: 10.1016/j.dss.2026.114653.
74. S. Bahadori and A. Hassanzadeh, "A comprehensive literature review on AI-based credit assessment in digital lending: Exploring models, transparency, alternative data, and real-time learning," Expert Systems with Applications, vol. 318, Art. no. 131872, 2026. doi: 10.1016/j.eswa.2026.131872.
75. S. Fritz-Morgenthal, B. Hein, and J. Papenbrock, "Financial risk management and explainable, trustworthy, responsible AI," Frontiers in Artificial Intelligence, vol. 5, Art. no. 779799, 2022. doi: 10.3389/frai.2022.779799.
76. F. Yang, M. Z. Abedin, and P. Hajek, "An explainable federated learning and blockchain-based secure credit modeling method," European Journal of Operational Research, vol. 317, no. 2, pp. 449–467, 2024. doi: 10.1016/j.ejor.2023.08.040.
77. Y. Gao, S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao, "Going beyond XAI: A systematic survey for explanation-guided learning," ACM Computing Surveys, vol. 56, no. 7, Art. no. 188, 2024. doi: 10.1145/3644073.
78. S. Hué, C. Hurlin, C. Pérignon, and S. Saurin, "Measuring the driving forces of predictive performance: Application to credit scoring," Management Science, published online Apr. 8, 2026. doi: 10.1287/mnsc.2023.02025.

Appendix A. Supplementary Reviewer Diagnostics

Table S1. LightGBM loan-to-value/property component ablations (exploratory)

scenario	removed	roc_auc	pr_auc	auc_change_from_full	pr_change_from_full
full_saved_reference	none	0.8654	0.7566	0	0
minus_ltv_value_and_status	loan_to_value_ratio_num, loan_to_value_ratio_status	0.8582	0.7429	-0.0072	-0.0137
minus_property_value_and_status	property_value_num, property_value_status	0.8578	0.7399	-0.0076	-0.0167
minus_both_numeric_values	loan_to_value_ratio_num, property_value_num	0.8557	0.7399	-0.0097	-0.0166
minus_both_explicit_statuses	loan_to_value_ratio_status, property_value_status	0.8632	0.7481	-0.0021	-0.0085
minus_both_complete_fields	loan_to_value_ratio_num, loan_to_value_ratio_status, property_value_num, property_value_status	0.8409	0.7082	-0.0245	-0.0484

Table S2. LightGBM income/debt-to-income component ablations (exploratory)

scenario	removed_features	roc_auc	pr_auc	delta_roc_auc	delta_pr_auc
saved_full_reference	none	0.8654	0.7566	0	0
remove_income_numeric	income_num	0.8620	0.7495	-0.0034	-0.0070
remove_income_status	income_status	0.8654	0.7566	1.39e-05	4.14e-06
remove_income_complete	income_num, income_status	0.8615	0.7487	-0.0038	-0.0079
remove_dti_complete	debt_to_income_ratio_public	0.8220	0.6612	-0.0433	-0.0954
remove_income_and_dti	debt_to_income_ratio_public, income_num, income_status	0.7934	0.6129	-0.0719	-0.1437

Table S3. leave-one-dimension-out technical means and ranks

Removed	Model	Score	Rank
NONE	Logistic Regression	49.8161	4
NONE	Random Forest	64.3620	1
NONE	XGBoost	63.2088	2
NONE	LightGBM	59.2543	3
Predictive Performance	Logistic Regression	55.7263	4
Predictive Performance	Random Forest	64.8560	1
Predictive Performance	XGBoost	62.1954	2
Predictive Performance	LightGBM	57.2422	3
Explainability	Logistic Regression	39.8552	4
Explainability	Random Forest	59.5963	1
Explainability	XGBoost	58.4344	2
Explainability	LightGBM	56.2021	3
Fairness	Logistic Regression	60.8678	4
Fairness	Random Forest	76.4102	2
Fairness	XGBoost	76.4221	1
Fairness	LightGBM	71.1803	3
Stability	Logistic Regression	49.4437	4
Stability	Random Forest	55.7764	1
Stability	XGBoost	54.5035	2
Stability	LightGBM	49.6287	3
Robustness	Logistic Regression	43.1877	4
Robustness	Random Forest	65.1710	1
Robustness	XGBoost	64.4886	2
Robustness	LightGBM	62.0181	3



Figure S1. Exploratory AUC changes after removing four underwriting fields.



Figure S2. Retrospective field missingness by recorded outcome.

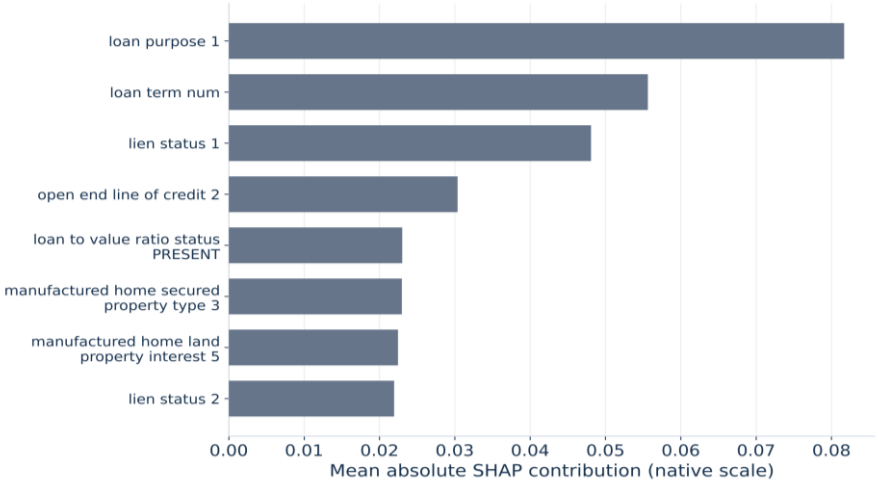


Figure S3. Logistic Regression SHAP attribution on 300 validation observations.

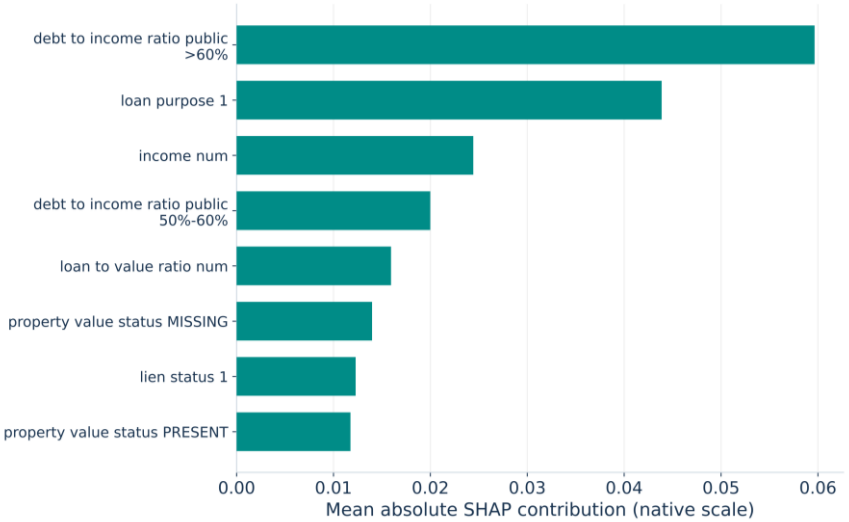


Figure S4. Random Forest SHAP attribution on 300 validation observations.

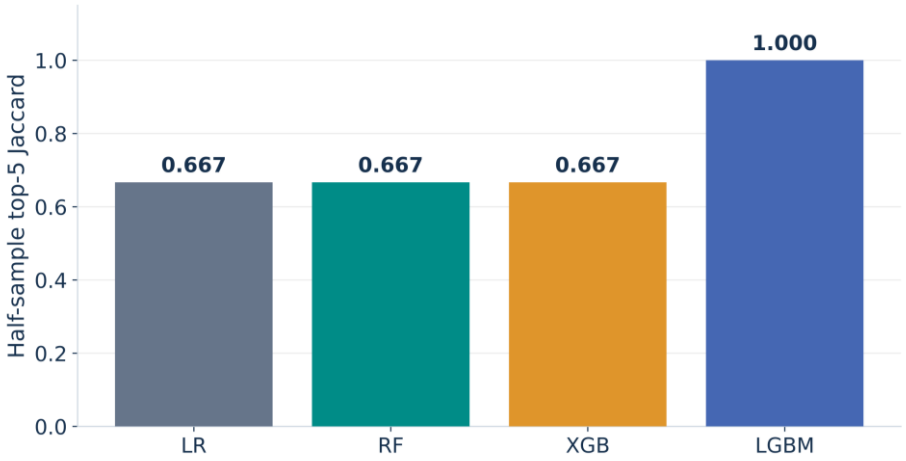


Figure S5. Half-sample top-five SHAP feature-ranking agreement.