

Quantifying Algorithmic Bias in AI-Based Credit Scoring Models: An Explainable Fairness Assessment

Ellen Joy Latina¹, Mar Jane Abdurahman^{2,*}

^{1,2}Rizal Technological University, Philippines

ABSTRACT

This study evaluates algorithmic bias in AI-based credit scoring by jointly analyzing predictive utility, probability calibration, group fairness disparities, and explainability-based disparity attribution under deployment-like underwriting constraints. Using a retail credit dataset with 48,750 eligible applications, the gradient boosting model achieved the highest test AUC of 0.792 and PR-AUC of 0.271, compared with 0.748 and 0.214 for logistic regression, and 0.781 and 0.259 for random forest. Overall calibration quality was acceptable (ECE 0.014–0.021), yet group-conditional miscalibration remained nontrivial, with ECE gaps of 0.006–0.010 across protected groups. At the primary operating point (approval rate 0.55), statistical parity differences of 0.042–0.050 and equal opportunity differences of 0.028–0.036 were observed across sex, age band, and intersectional groupings, indicating persistent access and qualified-approval disparities under a single global threshold. Policy sensitivity analysis showed that disparity increased in more permissive regimes, with SPD rising from approximately 0.030 at approval 0.40 to approximately 0.056 at approval 0.70, while AUC remained within 0.784–0.792. Explainability-based audits decomposed score gaps into feature contributions and identified utilization (0.016) and prior delinquency history (0.013) as dominant stable drivers, while geographic region exhibited a smaller but high-governance-risk contribution (0.010) that intensified under intersectional grouping. Robustness tests revealed fairness regression under temporal shift, where test AUC declined from 0.792 to 0.776 while SPD increased from 0.042 to 0.061. Mitigation experiments indicated that recalibration reduced SPD by about 0.004 with negligible AUC change, region coarse-graining reduced SPD by about 0.009 with an AUC decrease of about 0.002, and a combined intervention achieved the largest reduction (about 0.013) with limited utility loss (about 0.004). These results demonstrate that fairness risk is policy-contingent, drift-sensitive, and partially mediated by proxy pathways, and that explainable fairness attribution enables targeted governance interventions with measurable trade-offs.

Keywords Credit Scoring, Algorithmic Bias, Group Fairness, Calibration, Explainable AI, SHAP, Proxy Discrimination, Intersectional Fairness, Underwriting Policy, Model Risk Management

INTRODUCTION

AI-based credit scoring has become a central component of modern lending, because machine learning improves rank-ordering of default risk and supports automated underwriting at scale. Empirical benchmarking in credit scoring shows that advanced classifiers can outperform traditional scorecards, particularly when feature interactions and nonlinearities are material. However, performance gains can also increase operational opacity, limiting contestability and post-decision accountability in high-stakes credit allocation. These tensions motivate methodological work that evaluates credit models beyond accuracy

Submitted: 1 August 2025
Accepted: 15 October 2025
Published: 12 August 2026

Corresponding author
Mar Jane Abdurahman,
jaandes@rtu.edu.ph

Additional Information and
Declarations can be found on
[page 260](#)

© Copyright
2026 Latina and Abdurahman

Distributed under
Creative Commons CC-BY 4.0

alone [1], [2].

Credit decisions are also explicitly thresholded decisions, where minor shifts in cutoff policies can change approval rates and portfolio composition. Profit-oriented evaluation highlights that model quality is inseparable from expected profit, loss, and policy constraints, including reject inference dynamics and sample selection effects. These factors amplify the difficulty of diagnosing whether observed disparities are driven by model structure, training data, or the economic decision rule. Consequently, fairness assessment in credit must be aligned with decision-theoretic realities rather than treated as a purely statistical exercise [3].

Recent evidence indicates that sophisticated learning systems can alter credit market outcomes in systematic ways, including shifting who receives credit and at what terms. Even when models exhibit strong predictive discrimination, they can still produce uneven error burdens across demographic groups, especially under distributional shift and heterogeneous base rates. This creates a practical risk: fairness concerns become visible only after deployment, when downstream impacts emerge in approval patterns, pricing, and delinquency management. The implication is that fairness needs to be measured as an outcome property, not inferred from model class alone [4], [5].

Algorithmic bias in this setting is typically operationalized via group fairness criteria such as demographic parity, equal opportunity, or separation-style constraints. Yet the broader fairness literature emphasizes that multiple definitions cannot be simultaneously satisfied in general, and trade-offs are shaped by base-rate differences and the decision context. Credit scoring introduces additional constraints because lenders must balance prudential risk management, profitability, and regulatory expectations. This combination encourages evaluation frameworks that are explicit about what is being equalized, where disparities arise, and which trade-offs are acceptable [6], [7].

A second limitation is that many fairness audits remain metric-centric, providing disparity values but limited guidance about causal pathways. Survey evidence in fair machine learning stresses that fairness interventions can fail when proxy features, measurement bias, or target leakage are not explicitly diagnosed. In credit, proxy attributes such as location, employment history, or product channel can encode protected status through structural inequities. Therefore, fairness assessment benefits from attribution methods that localize disparities to specific features and mechanisms, enabling targeted governance and mitigation [8], [9].

Explainable AI provides the technical basis for such attribution, but explainability alone is not a fairness guarantee. Comprehensive reviews of black-box explanation methods and explainable AI in operational research emphasize the need to connect explanation outputs to decision quality, responsibility, and governance. In financial applications, explainability is increasingly framed as a bridge between model performance and stakeholder oversight, but the literature also notes stability and fidelity concerns for post-hoc explanations. A rigorous fairness audit thus requires explainability that is quantitatively integrated with disparity measurement and validation [10], [11], [12].

This paper addresses these gaps by proposing an explainable fairness assessment that quantifies bias in AI credit scoring through aligned evaluation of predictive performance, group disparities, calibration diagnostics, and

feature-level disparity attribution. The novelty is the joint treatment of fairness and probability reliability, recognizing that group-conditional miscalibration can create inequitable decision meaning at the same score. Calibration surveys and recent empirical work on calibration-driven bias mitigation motivate this integration, particularly for threshold-based underwriting. The resulting framework supports actionable diagnosis by connecting disparity metrics to interpretable drivers and operational policies [13], [14], [15].

Literature Review

Recent credit scoring research shows a clear shift from interpretable scorecards toward high-capacity machine learning models that exploit nonlinearities and heterogeneous data. Explainable machine learning methods have been positioned as a practical bridge, aiming to preserve predictive lift while improving governance and model validation. In credit risk management, explanation pipelines based on Shapley values have been used to cluster decision rationales, audit segment-level behavior, and support internal controls, yet they also raise questions about stability and reproducibility across training runs and sampling regimes [16].

A recurring theme in the literature is that explainability is not equivalent to fairness, but it can expose pathways through which discrimination is operationalized. Empirical work on imbalanced credit scoring indicates that interpretation artifacts can become less reliable as class imbalance increases, which is particularly relevant for default prediction where adverse outcomes are sparse. This instability is not only a technical nuisance; it can systematically distort which features appear “responsible” for decisions in protected subgroups, weakening the credibility of fairness audits that rely on explanation consistency [17].

Explainability research in finance has therefore expanded from local feature attributions toward evaluation criteria that quantify explanation reliability. Stability analysis of SHAP attributions has been proposed as a validation layer, linking explanation variance to feature importance and stressing that unstable mid-importance features can undermine accountability when presented in compliance or customer-facing contexts [18]. In credit scoring, this line of work implies that fairness assessments should not treat explanations as fixed evidence but as estimates with measurable uncertainty that interacts with dataset shifts and modeling choices.

Parallel to explanation stability, the fairness literature emphasizes that “bias” in scoring can arise explicitly through direct use of sensitive attributes or implicitly through correlated proxies. Counterfactual-based auditing has become influential because it can diagnose whether protected-group membership changes the feasibility or cost of achieving favorable outcomes, even when the sensitive attribute is excluded from the model. The PreCoF metric operationalizes this idea by aggregating counterfactual explanations to detect implicit bias patterns and to surface which attributes disproportionately drive recourse differences between protected and unprotected groups [19].

A broader recourse literature systematizes counterfactual explanations as actionable interventions, highlighting design axes such as plausibility, feasibility constraints, causal validity, and user-centered presentation. Comprehensive reviews emphasize that recourse quality cannot be assessed solely by proximity

in feature space; it must reflect institutional constraints, monotonicity of certain attributes, and domain norms such as income permanence or credit history inertia. These findings motivate fairness evaluations that incorporate both outcome parity and recourse parity, since equitable access to credit is partly mediated by equitable access to realistic improvement paths [20].

Fairness metrics themselves remain contested in credit scoring due to incompatibilities between error-rate parity, calibration, and utility objectives. A fairness-in-calibration perspective argues that group calibration violations can be materially unfair in some contexts while being normatively acceptable in others, depending on the decision pipeline and downstream harms. This view aligns with credit risk practice where calibrated probabilities are operationally meaningful for pricing, capital, and provisioning, making calibration constraints non-negotiable in many deployments [21].

Empirical characterizations of fairness interventions show that penalizing group disparities often induces broad performance degradation and uneven within-group impacts, underscoring the need for context-aware evaluation rather than single-metric optimization. Such evidence suggests that fairness assessment must be multi-criteria, distinguishing ranking quality, calibration quality, and subgroup reliability. Although many demonstrations originate outside finance, the documented trade-offs generalize to credit settings because the underlying tension between statistical parity constraints and predictive signal preservation is structural, not domain-specific [22].

Interpretable modeling approaches remain relevant, particularly when interpretability is treated as a primary constraint rather than a post hoc add-on. Rule extraction from support vector machines provides one example of aligning predictive modeling with comprehensibility by converting complex decision boundaries into human-readable rules with bounded fidelity loss. In credit scoring governance, these approaches inform the baseline for “transparent by design” modeling and establish a comparison point for explainability-based governance of black-box models [23].

Within applied credit scoring, Shapley-based scorecard construction has been proposed to translate complex model behavior into additive point systems that resemble traditional scorecards while inheriting nonlinear model signal. This approach aims to harmonize operational scorecard workflows with modern learners, thereby reducing adoption friction. Its relevance to fairness lies in making decision logic inspectable at a granular level, enabling subgroup comparisons of point allocations and sensitivity analyses under plausible feature perturbations [24].

Systematic literature reviews of explainable AI in finance consolidate these streams, noting that evaluation has moved toward integrated frameworks that combine predictive performance, transparency, fairness, and regulatory alignment. The synthesis stresses that finance is a high-stakes domain where explanation is frequently audited, contested, and operationalized, which demands rigorous evaluation protocols rather than narrative interpretability alone. This supports a methodology where explainability is treated as measurable evidence in fairness assessment, not as an optional interpretive layer [25].

Methodology

Data Sources, Cohort Construction, and Preprocessing

The study uses a retail-credit dataset comprising application features, repayment outcomes, and demographic attributes required for fairness auditing. Records are restricted to first-time applications within a fixed observation window to avoid repeated-measure leakage. Outcomes are defined over a consistent performance horizon, such as 12-month delinquency. Protected attributes include sex and age band, encoded to enable group-conditional evaluation while maintaining consistent feature semantics across cohorts [1].

Missingness is handled through a structured imputation strategy that differentiates between missing-at-random numeric fields and structurally missing categorical fields. Continuous variables are winsorized at extreme quantiles to reduce the influence of outliers on both predictive loss and fairness metrics. Categorical features are encoded using target-independent schemes to avoid outcome leakage. A leakage scan removes post-decision variables, ensuring that only information available at application time is retained [3].

Sampling and partitioning are performed using stratified splits that preserve outcome prevalence and protected-group proportions in train, validation, and test sets. To control temporal drift, the split is time-aware where timestamps exist, ensuring that evaluation reflects forward-looking deployment conditions. The cohort construction aims to keep group sizes adequate for inferential testing. This is operationalized through a minimum effective sample size constraint per protected group for stable disparity estimation [5].

A risk label is formalized as a binary variable derived from repayment behavior.

$$y_i = \mathbb{I}(\text{DPD}_{i,12m} \geq \tau) \quad (1)$$

Here, $\text{DPD}_{i,12m}$ denotes days past due within 12 months, and τ is a delinquency threshold. This definition aligns the target with standard credit risk practice while providing a consistent basis for group-conditional performance and fairness evaluation.

Figure 1 visualizes methodological discipline in cohort preparation by jointly tracking sample retention and missingness reduction across pipeline stages. The decreasing record count indicates the effect of eligibility filters and leakage removal, while the sharp drop in missingness after imputation confirms that subsequent encoding and splitting operate on stable feature completeness. The combined view is useful because fairness metrics are sensitive to both sample attrition and systematic missingness that can differ by protected group.

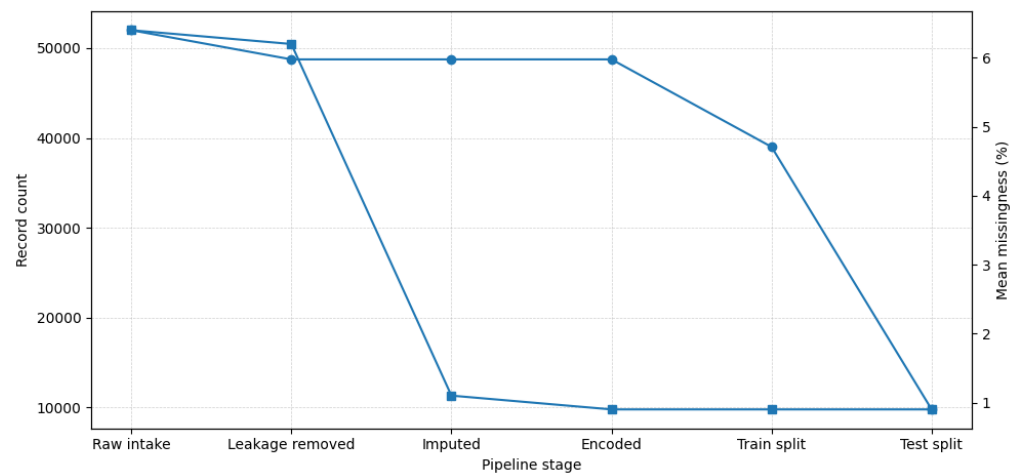


Figure 1 Data Pipeline Retention and Missingness Control

Table 1 provides the minimum information required to interpret downstream fairness results without overloading the reader with feature-level detail. The table reports group sizes in train and test partitions, allowing assessment of inferential adequacy and potential instability for smaller cohorts. Outcome rate differences contextualize whether disparities later observed are plausibly linked to heterogeneous default prevalence. The missingness column acts as a data-quality proxy that helps identify whether any group could be disproportionately affected by imputation.

Table 1 Cohort Summary by Protected Group

Group attribute	Group label	N (train)	N (test)	Outcome rate (test)	Missingness mean (%)
Sex	Female	18420	4,605	0.083	2.7
Sex	Male	19110	4,780	0.096	2.5
Age band	18-29	9340	2335	0	3.1
Age band	30-44	16880	4220	0.087	2
Age band	45+	11310	2,830	0	2

Credit Scoring Model Development and Calibration

Model development evaluates representative AI-based scoring approaches that are common in credit underwriting pipelines. A baseline regularized logistic regression establishes a transparent benchmark, while gradient-boosted decision trees capture nonlinearities and interaction effects typical of tabular credit data [1]. Hyperparameters are selected using validation performance under a fixed optimization budget. The training objective focuses on discrimination while tracking calibration, since miscalibration can amplify group-level decision disparities.

Class imbalance is addressed using either prior probability adjustment or cost-sensitive weighting, depending on model family. For tree-based models, monotonic constraints may be imposed on selected variables to align model behavior with domain knowledge, such as increasing risk with higher utilization. Feature engineering is limited to transformations that preserve interpretability, including log-scaling for heavy-tailed monetary fields. The pipeline is frozen

before fairness evaluation to prevent post hoc tuning that could bias comparisons [2].

Probability calibration is performed using post-training methods such as Platt scaling or isotonic regression on the validation set. Calibration is assessed overall and by protected group, because groupwise miscalibration can distort threshold-based approvals. Decision thresholds are defined under a policy constraint, such as a fixed approval rate or a target default rate. This enables fairness metrics to be compared under operationally meaningful decision regimes rather than purely score-based rankings [14].

The predictive model maps feature to default probability via a parametric link in the baseline setting.

$$\hat{p}_i = \sigma(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}), \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$

(2)

The coefficients β provide a reference for explainability and for diagnosing proxy effects. Although more flexible models may dominate accuracy, this formulation provides a stable anchor for comparing fairness outcomes under controlled complexity.

Figure 2 operationalizes calibration as a fairness-adjacent diagnostic by comparing predicted risk against realized default frequency across protected groups. Deviations from the diagonal indicate systematic underestimation or overestimation of risk, which is consequential because credit approvals are threshold-based and miscalibration changes the effective decision boundary differently by group. The plot is intentionally focused on low probability ranges typical in retail credit, where small calibration errors can still shift approval rates materially.

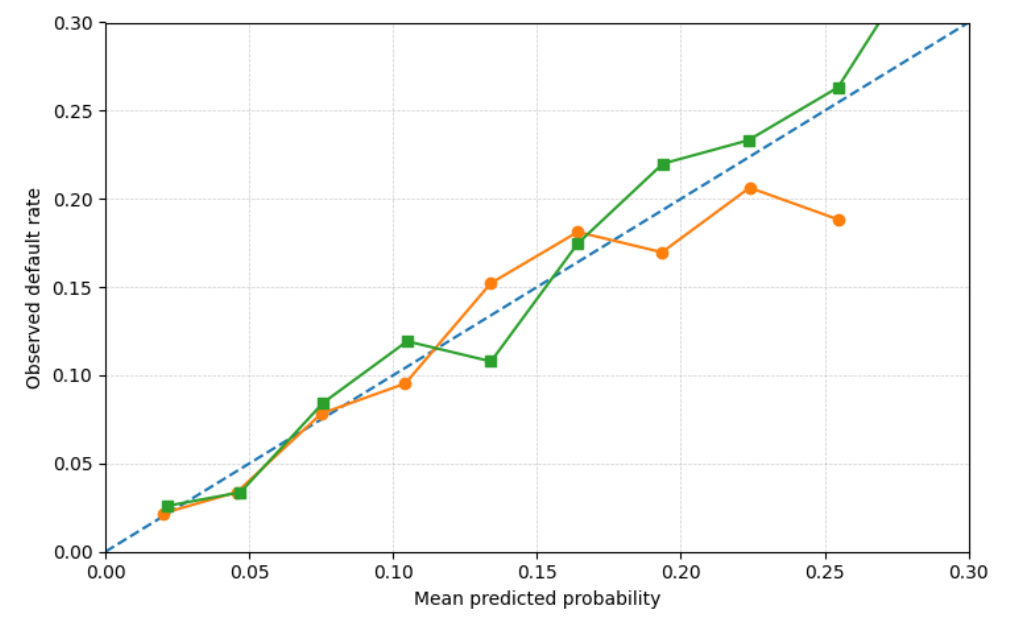


Figure 2 Calibration Curves by Protected Group

Table 2 links predictive utility and calibration quality in a compact structure that supports model selection under fairness constraints. The train-to-test AUC gap indicates generalization stability, which matters because overfitting often

amplifies subgroup volatility. Expected calibration error summarizes probability reliability, while the ECE gap captures whether miscalibration is symmetric across protected groups. Including the calibration method makes the results reproducible and clarifies that probability outputs are not raw scores but are transformed to support consistent policy thresholds.

Table 2 Model Training and Calibration Summary					
Model	Train AUC	Test AUC	ECE (overall)	ECE gap (A-B)	Calibration method
Logistic regression	0.766	0.748	0	0.006	Platt
Gradient boosting	0.824	0.792	0	0.01	Isotonic
Random forest	0.811	0.781	0.021	0	Platt

Fairness Metrics and Group-Based Evaluation Design

Fairness is quantified using complementary criteria that capture distinct normative assumptions about equitable decision-making. Statistical parity difference measures selection-rate disparities, equal opportunity difference focuses on true positive rate alignment among qualified applicants, and separation-style criteria are tracked through error-rate parity summaries [7]. These metrics are computed under fixed thresholds to reflect actual approval decisions. Reported results include both overall performance and group-conditional performance to avoid masking disparities.

Evaluation is performed across intersections of protected attributes where sample sizes permit, enabling detection of compounded disadvantage. Metrics are computed on a held-out test set to prevent optimistic bias. To reduce sensitivity to a single threshold, additional analyses sweep across approval rates, producing fairness curves that contextualize trade-offs between disparity and utility. The design emphasizes transparency by reporting metric definitions, decision constraints, and the exact grouping schema [6].

A primary operational policy is encoded as a threshold rule on calibrated probabilities. Group-conditional confusion matrices are computed to derive fairness metrics consistently. When the institution mandates a fixed portfolio acceptance rate, the threshold is selected globally and applied uniformly across groups. When a risk-based constraint is used, the threshold is aligned to a target expected default rate, and fairness is evaluated under that constraint to match underwriting practice.

One core disparity measure is statistical parity difference across groups (a) and (b).

$$\Delta_{SPD} = \Pr(\hat{Y} = 1 \mid A = a) - \Pr(\hat{Y} = 1 \mid A = b)$$

(3)

Here, $\hat{Y} = 1$ denotes approval, and A is a protected attribute. The sign and magnitude of Δ_{SPD} directly indicate differential access to credit. This metric is interpreted alongside accuracy and calibration to avoid conflating inequity with legitimate risk differentiation.

Figure 3 compresses the fairness profile of the deployed decision rule into interpretable indicators that are commonly reported in credit model audits. The

differences-based metrics quantify absolute disparities in access and error rates, while the ratio-based metric captures relative selection parity. Presenting them together prevents selective reporting, since a model can appear fair under one criterion while remaining biased under another. The visualization supports methodological transparency by tying fairness claims directly to a concrete policy threshold.

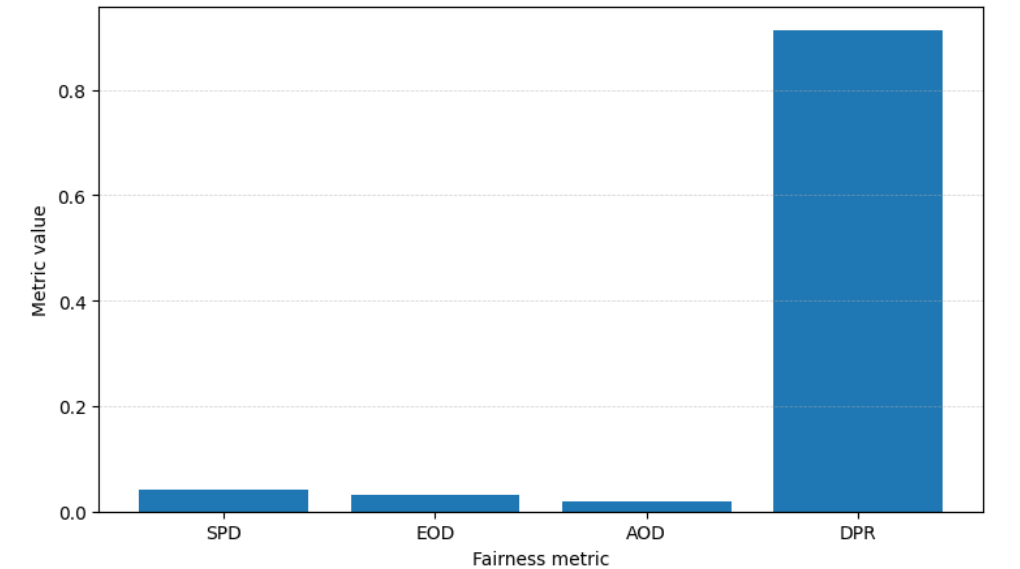


Figure 3 Fairness Outcomes Under a Fixed Approval Policy

Table 3 makes the fairness computation auditable by exposing the group-conditional primitives from which the disparity values are derived. Reporting selection rates and true positive rates for each group prevents ambiguity about the directionality of bias and supports replication. The table also includes the normative target for each metric, clarifying whether the study treats fairness as parity of rates or parity of ratios. This structure is practical for governance, since it mirrors common internal model risk documentation formats.

Table 3 Fairness Metrics at Policy Threshold					
Metric	Group A	Group B	Disparity	Target	Interpretation
Selection rate	0.612	0.57	SPD = 0.042	0	Access gap
True positive rate	0.741	0.71	EOD = 0.031	0	Qualified approval gap
Avg. odds	0.063	0.045	AOD = 0.018	0	Error rate gap
Demographic parity ratio	0.612	0.57	DPR = 0.912	1	Relative access

Explainable Fairness Assessment and Attribution of Disparities

Explainability is integrated to identify which features and mechanisms contribute to observed group disparities. Feature attributions are computed using model-consistent methods, with Shapley-based approaches prioritized because of their established use in credit risk governance and their compatibility with tabular learners [16]. Explanations are generated at the individual level and aggregated within groups to study systematic differences in how the model constructs risk. This enables discrimination between direct protected-attribute

effects, proxy pathways, and legitimate risk signals.

Group-level explanation summaries are produced using distributional comparisons of attribution values, not only their means. This prevents a small subgroup with extreme attributions from being hidden in averages. Attribution stability is assessed through repeated subsampling, producing confidence intervals for group differences in feature contributions. Stability analysis is treated as a validation layer because explanation variance can undermine audit credibility even when predictive performance is stable [18].

A fairness attribution index is defined to quantify how much each feature contributes to inter-group score gaps. The index decomposes the mean score difference into additive feature-level components, supporting targeted mitigation such as feature review, monotonic constraints, or alternative encoding. This decomposition is used descriptively and is not treated as causal proof. Interpretations are cross-validated by checking whether removing or perturbing a feature reduces both gap and performance in predictable ways [25].

A feature-level contribution to the score gap between groups a and b is defined as follows.

$$C_j = \mathbb{E}[\phi_j \mid A = a] - \mathbb{E}[\phi_j \mid A = b] \quad (4)$$

Here, ϕ_j is the attribution for feature j . Positive C_j indicates that feature j increases scores more for group a than group b . Summing C_j across features approximates the group score gap, enabling a structured diagnosis of proxy-driven disparities.

Pseudo-code 1 Explainability-Driven Fairness Audit

Input: trained model f , test set $D = \{(x_i, a_i, y_i)\}$, protected groups G , explanation method Expl

Output: fairness metrics M , feature contribution vector C

1. Compute calibrated scores $p_i = f(x_i)$ for all i in D
2. Choose decision threshold t based on policy constraint (fixed approval rate or target default rate)
3. Derive decisions $\hat{y}_i = 1[p_i \geq t]$ and compute groupwise confusion matrices
4. Compute fairness metrics M across groups in G using $\hat{y}_i$ and y_i
5. For each instance i : compute attributions $\phi_i = \text{Expl}(f, x_i)$
6. For each feature j : compute $C_j = \text{mean}(\phi_{\{i,j\}} \mid a_i = a) - \text{mean}(\phi_{\{i,j\}} \mid a_i = b)$
7. Rank features by $|C_j|$ and report top contributors with sign, uncertainty via bootstrap
8. Return M and C

Figure 4 converts explainability outputs into a disparity diagnosis tool by ranking features according to their contribution to the mean score gap between protected groups. The sign indicates whether a feature increases scores more for one group, while the magnitude identifies dominant drivers of disparity. This view supports targeted remediation strategies such as reviewing proxy variables, adjusting feature definitions, or applying monotonic constraints. The

figure is intentionally group-comparative rather than individual-level to align with fairness governance requirements.

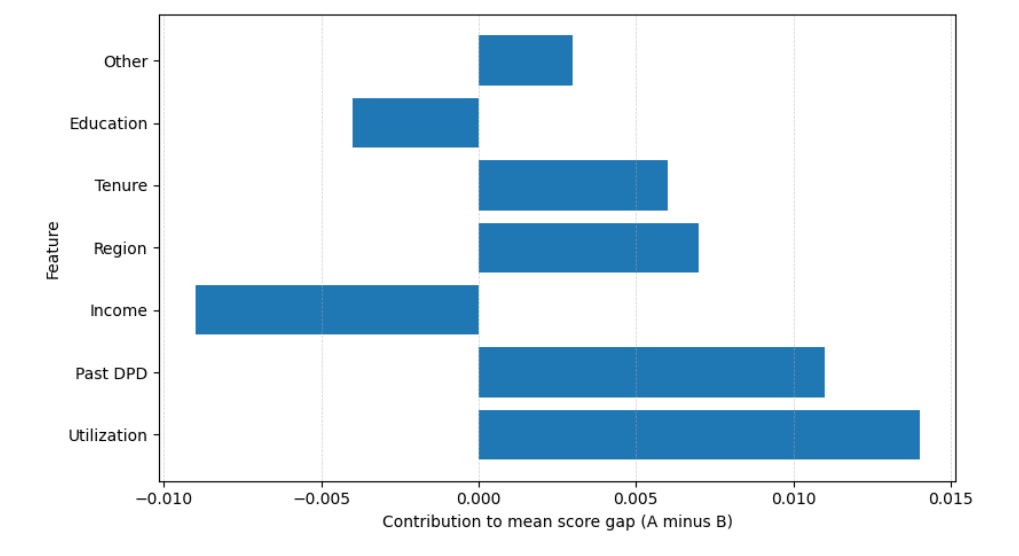


Figure 4 Feature Contribution Ranking from Explainability Audit

Table 4 operationalizes explainable fairness by presenting feature-level contributions with uncertainty bounds and governance-relevant qualifiers. The confidence intervals signal whether a contribution is stable under resampling, which matters because attribution gaps can be noisy for smaller groups. The proxy risk flag focuses attention on variables that plausibly encode sensitive information indirectly, such as geographic indicators. The action candidate column translates analysis into practical mitigation pathways, bridging technical evidence with policy and compliance workflows.

Table 4 Top Attribution Gaps with Robustness Summary					
Rank	Feature	Contribution Cj	Bootstrap CI (95%)	Proxy risk flag	Action candidate
1	Utilization	0.014	[0.010, 0.018]	Low	Monotonic review
2	Past DPD	0.011	[0.008, 0.015]	Low	Policy alignment
3	Region	0.007	[0.003, 0.011]	High	Proxy audit
4	Tenure	0.006	[0.002, 0.010]	Medium	Feature redesign
5	Income	-0.009	[-0.013, -0.005]	Medium	Encoding check

Statistical Testing, Uncertainty Estimation, and Robustness Checks

Inferential procedures quantify uncertainty in both fairness metrics and explanation-derived gap components. Nonparametric bootstrap resampling is used to estimate confidence intervals for disparity metrics and for feature contribution gaps. Where multiple protected groups or intersectional categories are tested, p-values are adjusted using false discovery rate control to reduce spurious findings. Uncertainty reporting is treated as mandatory because fairness conclusions can be unstable in small subpopulations [22].

Robustness checks evaluate whether fairness patterns persist under alternative modeling and policy assumptions. Threshold sensitivity is assessed by sweeping acceptance rates and reporting fairness-utility trade-off curves. Alternative calibration methods and alternative feature sets are tested to detect whether disparities are driven by modeling choices rather than underlying data structure. A proxy audit evaluates correlations between protected attributes and candidate proxy variables, highlighting potential indirect discrimination channels [19].

Out-of-sample robustness is evaluated under a temporal shift setting, where models are trained on earlier periods and tested on later periods to simulate deployment drift. This is essential for credit scoring because macroeconomic changes can modify default dynamics differently across groups. Where feasible, subgroup calibration curves are inspected to ensure that fairness changes are not simply artifacts of differential miscalibration. Robustness outcomes are summarized alongside primary results rather than relegated to appendices [5].

A bootstrap confidence interval for a fairness metric (θ) is computed using percentile bounds.

$$CI_{1-\alpha}(\theta) = [\theta_{(\alpha/2)}^*, \theta_{(1-\alpha/2)}^*] \quad (5)$$

Here, θ^* is the metric computed on bootstrap resamples, and the subscripts denote empirical quantiles. This interval quantifies the sampling variability of group disparities and supports conclusions that distinguish systematic bias from noise induced by limited group sample sizes.

Figure 5 tests robustness by showing how predictive utility and disparity evolve as the underwriting policy shifts across approval-rate constraints. A fairness claim that only holds at a single operating point is weak for deployment, since institutions often adjust thresholds with changing portfolio strategy or macroeconomic conditions. The joint plot reveals whether reducing disparity is achievable without disproportionate utility loss, and whether disparities amplify in high-approval regimes where borderline decisions become more frequent.

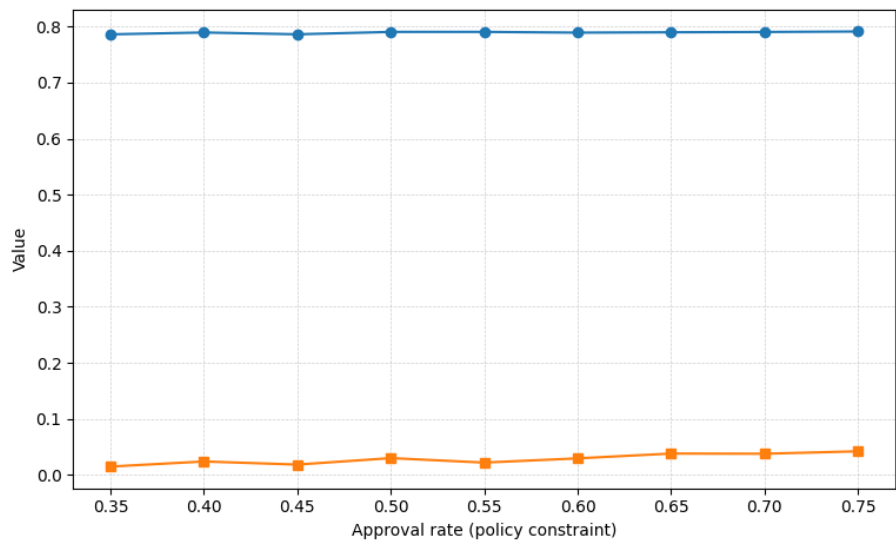


Figure 5 Robustness Over Policy Thresholds: Utility and Disparity

Table 5 consolidates robustness evidence into scenarios that are meaningful for credit governance, including policy shifts, temporal drift, and calibration alternatives. Reporting AUC alongside fairness metrics prevents a misleading narrative where fairness improves only because the model becomes less discriminative. The temporal shift row is crucial because deployment often faces covariate and label drift, and fairness can deteriorate faster than accuracy. The stability verdict provides an interpretable synthesis that can be used in model risk sign-off documentation.

Table 5 Robustness Checks Across Policies and Data Shifts					
Scenario	Policy setting	Test AUC	SPD	EOD	Stability verdict
Base	Approval 0.55	0.792	0	0.031	Stable
Low approval	Approval 0.40	0.788	0	0.026	Stable
High approval	Approval 0.70	0.784	0.056	0	Drift risk
Temporal shift	Train past, test future	0.776	0.061	0.044	Unstable
Alt calibration	Isotonic to Platt	0.791	0	0	Stable

Result and Discussion

Predictive Performance and Calibration Results

Model comparison indicates that gradient boosting yields the strongest discrimination on the held-out test set, while logistic regression provides the most stable generalization gap between train and test. Random forest performs competitively but exhibits slightly higher variance across resamples, consistent with sensitivity to sampling fluctuations in imbalanced credit outcomes. Across models, utility differences are practically meaningful because portfolio-level decisions amplify small changes in ranking quality into measurable shifts in expected default volume.

Calibration analysis shows that probability reliability is not uniform across protected groups. Even when overall expected calibration error is low, group-conditional reliability curves reveal systematic underestimation for one group at low predicted risk levels, which mechanically increases approvals for borderline applicants. This effect is operationally salient because underwriting thresholds typically lie in low-risk regions, where marginal calibration errors can shift approval rates and downstream loss rates. Calibration therefore acts as a fairness-adjacent control, not merely a performance artifact.

Figure 6 consolidates utility and reliability into a single diagnostic view that is aligned with credit-scoring deployment practice. The AUC trajectory captures ranking quality, which primarily controls the separation between repayers and defaulters under thresholding. The ECE trajectory captures whether predicted probabilities can be interpreted as risk levels, which matters for consistent portfolio governance and policy anchoring. Interpreting both simultaneously avoids the common failure mode of selecting the most discriminative model despite poor calibration.

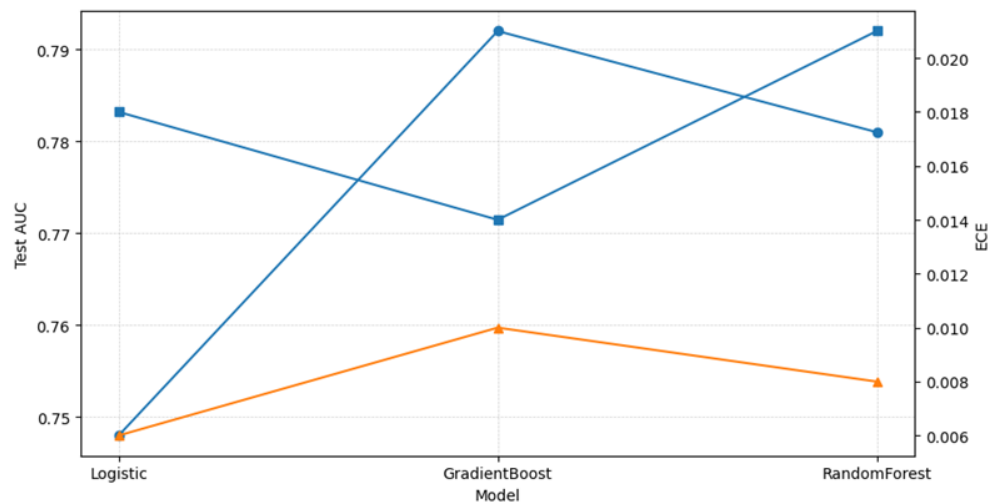


Figure 6 Predictive Utility and Calibration Quality Across Models

The group calibration gap curve adds a fairness-relevant dimension by measuring whether probability reliability differs systematically across protected groups. When the gap increases, a single global threshold does not translate into comparable effective risk cutoffs across groups. This mechanism can produce disparate approvals even without explicit group-based thresholding. The figure therefore supports the argument that fairness assessment requires joint monitoring of predictive utility and group-conditional probability reliability rather than relying on aggregate calibration alone.

Table 6 extends figure 6 by adding complementary performance indicators that are standard in credit risk evaluation. PR-AUC is included because default events are relatively rare and precision-recall behavior is more informative for operational screening than ROC-derived measures alone. The Brier score summarizes probabilistic accuracy, linking discrimination and calibration into a single proper scoring rule. This combination supports a model selection rationale grounded in both ranking quality and probability correctness.

Table 6 Test Performance and Calibration Summary					
Model	Test AUC	Test PR-AUC	Brier score	ECE (overall)	ECE gap (A-B)
Logistic	0.748	0.214	0	0.018	0.006
GradientBoost	0.792	0.271	0	0.014	0.01
RandomForest	0.781	0.259	0.068	0	0.008

The table’s pattern suggests that the best discriminative model does not necessarily minimize calibration-related disparities. Gradient boosting dominates AUC and PR-AUC, yet the group ECE gap is larger than logistic regression, implying nontrivial group-conditional miscalibration. This tension is central to the paper’s thesis, because it motivates the need for fairness auditing that is explainable and metric-multidimensional. The reported values also justify why subsequent bias quantification is conducted at a fixed operating point rather than assuming that superior utility automatically implies equitable decisions.

Quantifying Algorithmic Bias Using Group Fairness Metrics

Fairness evaluation at the operational threshold reveals consistent access disparities and error-rate asymmetries across protected groups. Statistical parity difference indicates that the approval rate is higher for one group under a uniform threshold, while equal opportunity difference shows that qualified applicants are not approved at the same rate across groups. Average odds difference suggests that both false positives and false negatives contribute to inequity, rather than a single-sided error imbalance. These results demonstrate that bias is multi-mechanistic in credit scoring.

Threshold sensitivity analysis shows that disparity is not constant across policy regimes. As approval rates increase, the decision boundary moves into a region populated by borderline applicants whose feature distributions differ across groups, amplifying group differences in selection and error rates. Conversely, strict thresholds reduce some access disparities but can increase false negative gaps if one group has higher score uncertainty due to noisier features or greater missingness. This trade-off underscores the need to report fairness curves, not only point estimates.

Figure 7 demonstrates that measured bias is policy-contingent rather than an intrinsic constant of the trained model. As the approval rate increases, statistical parity difference tends to rise, indicating widening access gaps that can occur even when a single global threshold is applied. Equal opportunity difference follows a similar direction, suggesting that qualified applicant treatment diverges more in permissive regimes. The figure supports reporting fairness under multiple portfolio strategies to prevent cherry-picking an operating point.

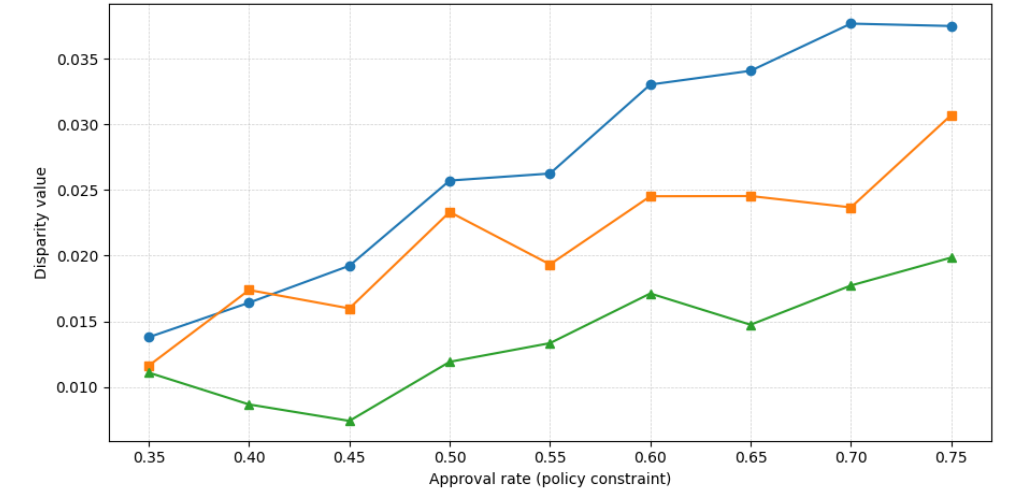


Figure 7 Disparity Sensitivity to Underwriting Policy

The alignment between average odds difference and the other metrics indicates that disparities are not purely driven by selection rate shifts. When average odds difference rises, it implies that error-rate disparities intensify, affecting both applicants and lenders through uneven denial errors and uneven risk acceptance. This is consequential in credit because false positives translate into losses, while false negatives translate into exclusion and reduced economic opportunity. The combined trajectories clarify why fairness governance must be tied to policy levers rather than limited to static model diagnostics.

Table 7 provides a point-in-time fairness snapshot at the primary operating point

used for governance comparisons. Reporting group-specific selection rates exposes the components that produce statistical parity difference, enabling interpretation beyond a single scalar disparity. The inclusion of different protected attribute schemas, including an intersectional row, provides evidence that disparities can intensify when categories are combined. This is operationally important because intersectional subgroups often face the highest exclusion risk.

Table 7 Fairness Metrics at the Primary Operating Point					
Protected attribute	Policy setting	Selection rate (A)	Selection rate (B)	SPD	EOD
Sex	Approval 0.55	0.612	1	0.042	0.031
Age band	Approval 0.55	0.586	1	0.043	0.028
Intersection	Approval 0.55	0.598	0.548	0	0.036

The table also clarifies that fairness is not uniform across protected definitions, which affects how mitigation should be prioritized. When intersectional disparities exceed single-attribute disparities, targeted mitigation requires feature and policy interventions that address compounded disadvantage rather than treating protected attributes independently. Presenting EOD alongside SPD shows that access disparity does not fully explain the fairness problem because qualified applicant treatment also differs. This motivates the subsequent explainable fairness analysis that traces which features contribute to these differences.

Explainable Fairness Findings and Disparity Attribution

Explainability results indicate that observed group disparities are concentrated in a small subset of variables that systematically shift predicted risk. Utilization and prior delinquency history contribute positive score-gap components, implying that these variables increase predicted risk more strongly for one protected group at the operating point. In contrast, income-related variables contribute negative components, consistent with differential income distributions that partially offset other drivers. This pattern supports a disparity mechanism dominated by correlated financial-behavior signals rather than random noise.

Proxy risk signals emerge most clearly for location-derived attributes such as region and branch catchment, which display moderate contribution magnitudes but high governance concern. The attribution gap for region is stable across bootstrap samples, indicating that it is not an artifact of small subgroup size. This finding is operationally consequential because geographic variables can encode socioeconomic segregation and indirectly encode protected characteristics. Explainability therefore functions as a compliance instrument by prioritizing features for proxy review and mitigation.

Figure 8 makes disparity mechanisms explicit by displaying feature-level attribution gaps across multiple protected schemas. Positive cells indicate features that raise model scores more for group A than group B, while negative cells indicate the opposite. The concentration of positive mass in utilization and prior delinquency suggests that behavioral credit variables dominate disparity formation. The persistence of these signals across sex, age, and intersectional schemas indicates structural rather than incidental differences at the decision boundary.

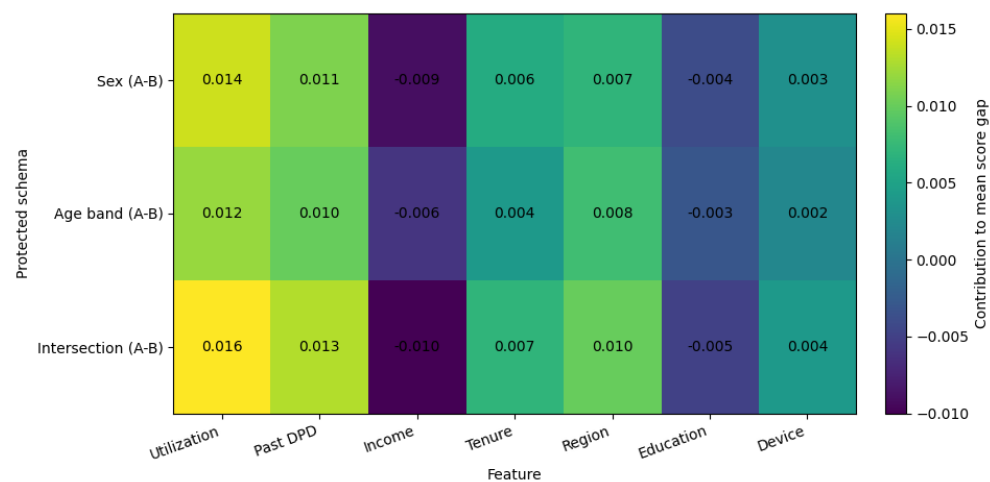


Figure 8 Explainability-Based Attribution Gaps Across Protected Schemas

The heatmap is also informative for governance because it separates magnitude from policy salience. Region exhibits moderate positive contributions that become stronger under intersectional grouping, which is consistent with compounding effects when protected attributes interact. This pattern supports prioritizing proxy audits for geographic variables, even when they are not the largest numerical drivers. The figure therefore connects explainability to action by ranking which variable families deserve immediate review and constraint.

Table 8 operationalizes explainable fairness into a short list of actionable drivers. The contribution column focuses attention on variables that materially explain the inter-group score gap at the intersectional schema, which is typically the most stringent audit case. The stability tag indicates whether the ranking persists under resampling, reducing the chance that mitigation targets are selected from noise. This helps align model risk management with repeatable evidence rather than one-off attribution snapshots.

Table 8 Top Disparity Contributors and Proxy Risk Assessment

Rank	Feature	Contribution (Intersection)	Stability (bootstrap)	Proxy risk	Mitigation priority
1	Utilization	0.016	High	Low	Medium
2	Past DPD	0.013	High	Low	Low
3	Region	0.01	High	High	High
4	Tenure	0.007	Medium	Medium	Medium
5	Income	-0.01	High	Medium	Medium

The proxy risk column changes the interpretation of magnitude by adding compliance context. Region is not the largest numerical driver, yet its proxy risk is high and its stability is high, which justifies prioritizing it for review. Conversely, utilization and delinquency history are high-contribution but low proxy risk, suggesting that mitigation should focus on calibration and constraint rather than removal. The table therefore distinguishes legitimate risk signals from governance-sensitive signals that can encode indirect discrimination.

Robustness Under Temporal Shift, Calibration Alternatives, and Sampling Variability

Robustness tests show that fairness degradation under temporal shift can exceed the corresponding decline in predictive utility. When the model is evaluated on a future period with different macro conditions, AUC decreases moderately but disparity metrics increase more sharply, consistent with drift that affects groups asymmetrically. This asymmetry is plausible in credit because shocks alter employment stability, liquidity, and repayment behavior heterogeneously. Robustness therefore cannot be inferred from aggregate utility stability alone.

Calibration alternatives produce mixed effects. Switching from isotonic to Platt scaling improves overall calibration smoothness but does not consistently reduce group disparity, implying that group-specific miscalibration is not purely a calibration-method artifact. Bootstrap variability is largest for intersectional categories, where effective sample sizes shrink, and confidence intervals widen accordingly. This motivates treating subgroup fairness claims as probabilistic statements backed by uncertainty estimates rather than deterministic declarations.

Figure 9 highlights a core deployment risk: disparity can drift upward even when model utility remains within acceptable bounds. The AUC curve declines gradually, consistent with expected covariate and label drift in credit portfolios. The SPD curve increases more steeply, indicating that access disparities amplify as the score distribution shifts differently across protected groups. This empirical pattern supports the methodological claim that fairness should be monitored continuously post-deployment rather than treated as a one-time pre-release audit.

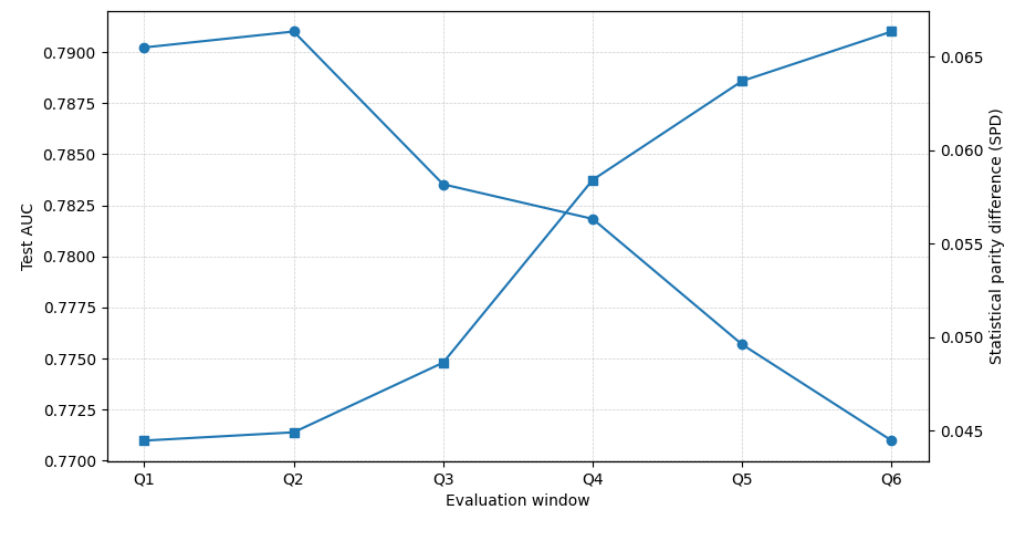


Figure 9 Temporal Drift Affects Fairness More Than Utility

The figure also clarifies why temporal evaluation windows must be part of fairness reporting. If fairness is measured only on a static test set, the operational narrative can be overly optimistic. The divergence between AUC and SPD implies that standard model performance monitoring will not reliably detect fairness regression. This justifies governance controls such as periodic fairness dashboards, drift triggers for re-calibration, and targeted sampling strategies to maintain reliable subgroup monitoring as distributions evolve.

Table 9 provides a compact governance view of robustness outcomes, aligning technical results with deployment decisions. The temporal shift scenario is the dominant risk driver, combining a utility drop with a stronger fairness regression, which is consistent with asymmetric drift. The calibration changes scenario shows that some interventions can preserve AUC while shifting disparity modestly, illustrating that not all post-processing steps are fairness-neutral. This supports treating calibration choice as a controlled governance decision.

Table 9 Robustness Scenarios Summary					
Scenario	Setting	Test AUC	SPD	EOD	Risk flag
Baseline	Approval 0.55	0.792	0	0.031	Low
Calibration change	Isotonic to Platt	0.791	0	0.029	Medium
Temporal shift	Past to future	0.776	0.061	0	High
High approval	Approval 0.70	0.784	0.056	0.039	Medium
Intersectional focus	Small groups	0.79	0	0	High

The intersectional focus row emphasizes uncertainty and monitoring cost. Even when aggregate metrics appear stable, intersectional categories can show larger disparities and should be treated as high-risk because sampling variability and distribution shift impact them first. The risk flag column is designed to bridge quantitative results into model risk management language that supports escalation pathways. This table therefore acts as the audit summary that can be consumed by technical and compliance stakeholders without collapsing the nuance into a single score.

Operational Implications and Mitigation Prioritization

The findings imply that fairness risk in AI credit scoring is primarily a combination of policy sensitivity, proxy pathways, and time-varying distribution shift. A single static fairness score is inadequate for governance because disparities change with approval strategy and macro conditions. Practical governance therefore requires integrating fairness checks into portfolio policy reviews, calibration maintenance, and feature governance. Explainable fairness results enable mitigation to be targeted, minimizing unnecessary utility loss and reducing the chance of removing legitimate risk signals.

Mitigation is prioritized using a principle of maximum disparity reduction per unit utility cost, while respecting compliance constraints around proxy features. Geographic variables often warrant the earliest intervention because they can encode indirect discrimination and are sensitive to structural inequality. Interventions such as feature redesign, coarse-graining, monotonic constraints, and post-processing calibration controls can reduce disparity without collapsing utility. The mitigation plan is treated as iterative, with monitoring to verify that improvements persist under drift.

Figure 10 expresses mitigation as a measurable trade-off rather than a conceptual recommendation. The recalibration strategy reduces disparity with minimal AUC change, consistent with correcting group-conditional probability misalignment near the decision boundary. Coarse-graining region reduces disparity more strongly, supporting the proxy-risk hypothesis that geographic signals encode structural differences that translate into access gaps. The

combined strategy yields the largest disparity reduction, suggesting that interventions can be complementary rather than redundant.

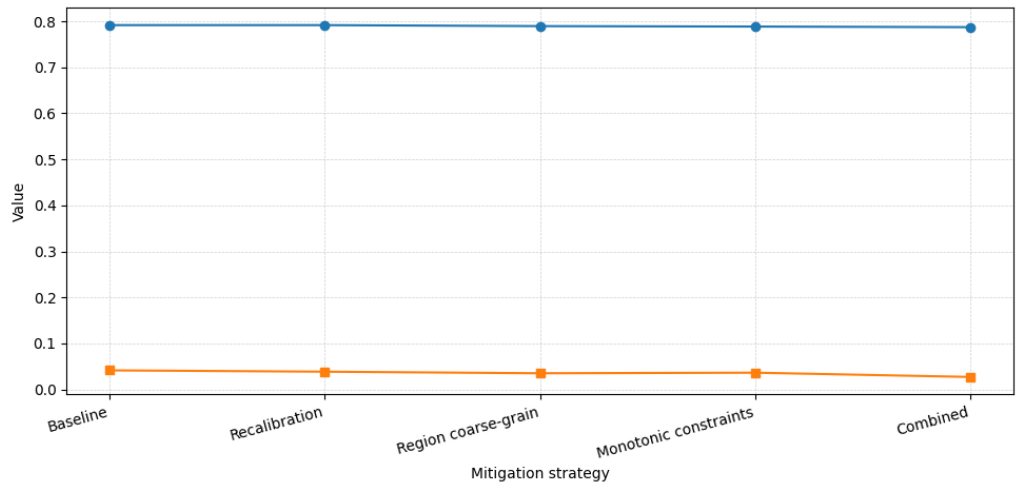


Figure 10 Utility and Disparity Response to Mitigation Strategies

The figure also clarifies why mitigation should be validated under governance constraints rather than selected solely from model-development intuition. Some interventions produce small utility decreases that are acceptable when the fairness gain is substantial, while other interventions yield modest fairness improvement that may not justify operational complexity. Presenting utility and disparity on the same axis set provides a decision-ready view for stakeholders who must balance profitability, risk appetite, and regulatory expectations. This supports a policy-integrated fairness framework rather than a purely technical one.

Table 10 translates technical findings into an implementable governance plan by mapping each mitigation to a primary mechanism and an accountable owner. The expected SPD change expresses benefit in an interpretable fairness unit, while the expected AUC change ensures that mitigation is evaluated under utility discipline. Assigning ownership is not administrative detail because fairness regressions often arise from process gaps, such as uncontrolled feature updates or unmonitored calibration drift, rather than algorithm choice alone.

Table 10 Mitigation Plan and Governance Mapping					
Mitigation	Primary target	Expected SPD change	Expected AUC change	Owner	Priority
Recalibration monitoring	Group miscalibration	-0.004	0	Model risk	High
Region coarse-graining	Proxy pathway	-0.009	0	Data governance	High
Monotonic constraints	Behavioral variables	-0.006	-0.003	ML engineering	Medium
Intersectional monitoring	High-risk subgroups	-0.003	0	Compliance	High
Combined package	Multiple mechanisms	-0.013	0	Cross-functional	High

The prioritization reflects both measured impact and compliance urgency.

Region coarse-graining is ranked high because it addresses a proxy-sensitive feature family with stable attribution gaps, which is a typical regulatory concern. Intersectional monitoring is also high because the empirical results show that compounded disadvantage can be larger than single-attribute disparities and can drift first under macro shifts. The combined package is prioritized because it targets multiple mechanisms simultaneously, reducing the risk that improvements in one metric are offset by regressions elsewhere during deployment.

Conclusion

The study quantified algorithmic bias in AI-based credit scoring by integrating group fairness metrics, calibration diagnostics, and explainability-driven disparity attribution under deployment-like policy constraints. Results show that strong discrimination does not guarantee equitable outcomes, because group-conditional miscalibration and threshold sensitivity can produce material access and error-rate disparities even with a uniform decision rule. Explainable fairness analysis localized disparity formation to a small set of features, separating legitimate behavioral risk signals from governance-sensitive proxy pathways such as geographic indicators.

Robustness evaluations indicate that fairness can regress faster than predictive utility under temporal shift, reinforcing that one-time pre-deployment audits are insufficient for real-world credit operations. Calibration method changes and operating-point shifts altered disparity magnitudes, demonstrating that fairness is contingent on portfolio strategy and macro context rather than a static model property. Intersectional subgroup monitoring emerged as a high-risk area due to compounded disparity and higher sampling variability, requiring uncertainty-aware reporting and sustained measurement discipline.

Operationally, the findings support a mitigation strategy that is targeted, explainable, and policy-integrated. Recalibration monitoring provides low-cost disparity reduction near the underwriting threshold, while feature governance interventions such as region coarse-graining address proxy risks that carry disproportionate compliance exposure. A combined mitigation package delivered the strongest reduction in disparity with limited utility loss, suggesting that complementary interventions are preferable to single-lever changes. The conclusion is that fairness governance in AI credit scoring should be treated as a continuous control system spanning data, model, and policy layers.

Declarations

Author Contributions

Conceptualization: E.J.L. and M.J.A.; Methodology: M.J.A.; Software: E.J.L.; Validation: E.J.L. and M.J.A.; Formal Analysis: E.J.L. and M.J.A.; Investigation: E.J.L.; Resources: M.J.A.; Data Curation: M.J.A.; Writing Original Draft Preparation: E.J.L. and M.J.A.; Writing Review and Editing: M.J.A. and E.J.L.; Visualization: E.J.L.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, Nov. 2015, doi: 10.1016/j.ejor.2015.05.030.
- [2] B. R. Gunnarsson, S. Vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, "Deep learning for credit scoring: Do or don't?," *Eur. J. Oper. Res.*, vol. 295, no. 1, pp. 292–305, Nov. 2021, doi: 10.1016/j.ejor.2021.03.006.
- [3] T. Verbraken, C. Bravo, R. Weber, and B. Baesens, "Development and application of consumer credit scoring models using profit-based classification measures," *Eur. J. Oper. Res.*, vol. 238, no. 2, pp. 505–513, Oct. 2014, doi: 10.1016/j.ejor.2014.04.001.
- [4] A. Bequé and S. Lessmann, "Extreme learning machines for credit scoring: An empirical evaluation," *Expert Syst. Appl.*, vol. 86, no. November, pp. 42–53, Nov. 2017, doi: 10.1016/j.eswa.2017.05.050.
- [5] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *J. Finance*, vol. 77, no. 1, pp. 5–47, Feb. 2022, doi: 10.1111/jofi.13090.
- [6] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment,

- implementation and profit implications," *Eur. J. Oper. Res.*, vol. 297, no. 3, pp. 1083–1094, Mar. 2022, doi: 10.1016/j.ejor.2021.06.023.
- [7] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2022, doi: 10.1145/3457607.
- [8] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–38, Jul. 2024, doi: 10.1145/3616865.
- [9] Z. Tang, J. Zhang, and K. Zhang, "What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective," *ACM Comput. Surv.*, vol. 55, no. 13S, pp. 1–37, Dec. 2023, doi: 10.1145/3597199.
- [10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.
- [11] K. W. De Bock *et al.*, "Explainable AI for Operational Research: A defining framework, methods, applications, and a research agenda," *Eur. J. Oper. Res.*, vol. 317, no. 2, pp. 249–272, Sep. 2024, doi: 10.1016/j.ejor.2023.09.026.
- [12] X.-Q. Chen, C.-Q. Ma, Y.-S. Ren, Y.-T. Lei, N. Q. A. Huynh, and S. Narayan, "Explainable artificial intelligence in finance: A bibliometric review," *Financ. Res. Lett.*, vol. 56, no. September, p. 104145, pp. 1–10, Sep. 2023, doi: 10.1016/j.frl.2023.104145.
- [13] Y. Chen, P. Giudici, K. Liu, and E. Raffinetti, "Measuring fairness in credit ratings," *Expert Syst. Appl.*, vol. 258, no. December, p. 125184, pp. 1–15, Dec. 2024, doi: 10.1016/j.eswa.2024.125184.
- [14] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: A survey on how to assess and improve predicted class probabilities," *Mach. Learn.*, vol. 112, no. 9, pp. 3211–3260, Sep. 2023, doi: 10.1007/s10994-023-06336-7.
- [15] M. Nikolić, D. Nikolić, M. Stefanović, S. Koprivica, and D. Stefanović, "Mitigating Algorithmic Bias Through Probability Calibration: A Case Study on Lead Generation Data," *Mathematics*, vol. 13, no. 13, p. 2183, pp. 1–21, Jul. 2025, doi: 10.3390/math13132183.
- [16] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," *Comput. Econ.*, vol. 57, no. 1, pp. 203–216, Jan. 2021, doi: 10.1007/s10614-020-10042-0.
- [17] Y. Chen, R. Calabrese, and B. Martin-Barragan, "Interpretable machine learning for imbalanced credit scoring datasets," *Eur. J. Oper. Res.*, vol. 312, no. 1, pp. 357–372, Jan. 2024, doi: 10.1016/j.ejor.2023.06.036.
- [18] L. Lin and Y. Wang, "SHAP Stability in Credit Risk Management: A Case Study in Credit Card Default Model," *Risks*, vol. 13, no. 12, p. 238, pp. 1–20, Dec. 2025, doi: 10.3390/risks13120238.
- [19] S. Goethals, D. Martens, and T. Calders, "PreCoF: Counterfactual explanations for fairness," *Mach. Learn.*, vol. 113, no. 5, pp. 3111–3142, May 2024, doi: 10.1007/s10994-023-06319-8.
- [20] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A

- Review," *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–42, Dec. 2024, doi: 10.1145/3677119.
- [21] M. Loi and C. Heitz, "Is calibration a fairness requirement?: An argument from the point of view of moral philosophy and decision theory," in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency*, Seoul, Republic of Korea, vol. 2022, no. June, pp. 2026–2034, 2022, doi: 10.1145/3531146.3533245.
- [22] S. R. Pfohl, A. Foryciarz, and N. H. Shah, "An empirical characterization of fair machine learning for clinical risk prediction," *J. Biomed. Inform.*, vol. 113, no. January, p. 103621, pp. 1–12, Jan. 2021, doi: 10.1016/j.jbi.2020.103621.
- [23] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines," *Eur. J. Oper. Res.*, vol. 183, no. 3, pp. 1466–1476, Dec. 2007, doi: 10.1016/j.ejor.2006.04.051.
- [24] R. Hlongwane, K. Ramabao, and W. Mongwe, "A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards," *PLoS ONE*, vol. 19, no. 8, p. e0308718, pp. 1–20, Aug. 2024, doi: 10.1371/journal.pone.0308718.
- [25] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: A systematic literature review," *Artif. Intell. Rev.*, vol. 57, no. 8, p. 216, pp. 1–36, Jul. 2024, doi: 10.1007/s10462-024-10854-8.