

Responsible Credit-Line Increase: A Fairness-Aware, Explainable, and Governed Machine-Learning Framework for Socio-Technical Lending Systems

Akshay Sharma

akshay.sharma1088@gmail.com

Salesforce (United States)

Gaurav Sharma

Institute of Electrical and Electronics Engineers

Research Article

Keywords: Responsible AI, Algorithmic fairness, Fair lending, Explainable AI, Probability calibration, Model risk management

Posted Date: October 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11530197/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Automated credit-line decisions shape access to liquidity for millions of consumers and are subject to fair-lending law. We study proactive credit-line increase (CLI) as a responsible-AI problem in which economic optimization, algorithmic fairness, explainability, and governance are designed together. We formalize CLI as incremental-profit maximization under a portfolio expected-loss budget and explicit fairness tolerances, and reduce it to a calibrated, auditable threshold policy with group-level recalibration that controls disparate impact without altering the validated model. Because account-level credit data cannot be shared, we evaluate on a released synthetic portfolio of 60,000 accounts. On a time-held-out fold, gradient boosting and a logistic baseline discriminate almost equally (AUC 0.762 vs. 0.758) and perform alike at a matched approval rate, while isotonic calibration of the boosted scores reduces expected calibration error from 0.194 to 0.006. The model is equally accurate across protected groups, yet a single threshold produces disparate impact (demographic-parity gap 0.084, disparate-impact ratio 0.85); group-wise recalibration under the same loss budget reduces the parity gap to 0.013 at a 0.6% cost in realized profit. A fairness-economics frontier shows that closing the parity gap by extending access to the disadvantaged group is profit-neutral to profit-positive in this portfolio, while a multi-period simulation shows that sustaining parity narrows the credit-access gap by 43% at a 0.6% cost in cumulative profit. On a real public credit benchmark the models again perform alike and the fairness controls transfer. We map each component to responsible-AI principles and model-risk governance, offering a reproducible reference design.

1 Introduction

Credit-limit management is a foundational mechanism in revolving credit portfolios. In a credit-card program, the assigned limit governs spending capacity and therefore shapes revenue, engagement, and default exposure simultaneously. A well-timed credit-line increase (CLI) can stimulate transaction volume, lift interchange and interest income, and strengthen retention; a poorly calibrated one elevates probability of default (PD), loss given default (LGD), and regulatory capital consumption. CLI is therefore not a pure prediction problem but a capital-allocation decision under uncertainty.

Historically, CLI decisions have been driven by static scorecards combined with periodic bureau refreshes, triggered by campaigns or manual review [1], [17], [33]. The central limitation is temporal: such approaches cannot continuously incorporate short-horizon dynamics in utilization, payment behavior, and behavioral volatility. Modern lakehouse data infrastructures now make near-real-time aggregation of transaction and payment signals practical, creating an opportunity to move from reactive review toward proactive, behavior-driven optimization.

CLI is also a socio-technical system, not merely a technical one. Access to revolving credit affects household liquidity, resilience to income shocks, and the ability to build credit history; the algorithm that decides who receives additional credit therefore redistributes financial opportunity at scale [39], [41]. Empirical studies of algorithmic lending show that machine-learning models can both reduce and

entrench disparities depending on how they are designed and governed [42], [43]. In the United States the decision is bound by fair-lending law: the Equal Credit Opportunity Act and Regulation B prohibit both disparate treatment and unjustified disparate impact, and the Fair Credit Reporting Act requires adverse-action explanations, so fairness and explainability are legal obligations rather than optional features. Treating CLI as a responsible-AI problem means designing for the borrower, the institution, the regulator, and society together.

Deploying machine learning for CLI thus raises four coupled requirements that the literature rarely addresses together. First, risk control: decisions must respect a portfolio-level expected-loss budget tied to risk appetite and capital. Second, fairness: outcomes must be monitored and controlled for disparate impact across protected groups [10]–[13], [26], [27]. Third, explainability: regulated lending requires transparent, auditable, contestable reasoning for each decision [6], [8], [28], [48]. Fourth, governance: models must align with model-risk-management expectations such as SR 11 – 7 [14] and with emerging responsible-AI principles [44], [45], including independent validation, ongoing monitoring, documentation, and human oversight. Most prior work optimizes classification accuracy in isolation and treats fairness, explainability, and governance as afterthoughts.

Research objective and contributions. This paper develops a fairness-aware, explainable, and governed CLI framework and evaluates it as a reproducible simulation study positioned within the responsible-AI and socio-technical-systems literature. Our contributions are: (i) a formal statement of proactive CLI as constrained incremental-profit maximization under a portfolio expected-loss budget and explicit fairness tolerances; (ii) a calibrated gradient-boosting PD pipeline whose probabilities are accurate enough to enforce the loss budget, with verified parity of accuracy across protected groups; (iii) a post-hoc, group-level threshold-recalibration mechanism that controls disparate impact and equalized-odds gaps without retraining or altering the validated model, together with an empirical fairness–economics frontier showing when equity and profitability align; (iv) an explicit mapping of the system to responsible-AI principles (fairness, transparency, accountability, contestability) and to model-risk and prudential governance; and (v) an end-to-end, fully released synthetic benchmark and code base that quantify the economic, fairness, and governance effects of each component.

2 Related work

2.1 Credit Scoring and Machine Learning

Logistic regression remains the regulatory reference model in credit scoring because of its additive, auditable structure [1], [2], [17], [33]; comprehensive overviews of consumer credit-risk assessment trace its long dominance and the gradual adoption of statistical learning [31], [32]. Its weakness is well documented: it assumes linearity in log-odds and captures feature interactions only when they are hand-engineered. Large benchmarking studies show that ensemble methods, and gradient-boosted trees in particular, consistently improve discrimination on tabular credit data [3], [18], [19], [24], [25], with deep

models offering limited additional gains on structured features [30], [40]. We adopt gradient boosting [4], [5], [24] for this reason, while retaining logistic regression as an interpretable baseline.

2.2 Explainability

Post-hoc attribution methods such as LIME [8] and SHAP [6], [7] approximate model reasoning with local, additive explanations, complementing broader treatments of interpretable machine learning [9]; SHAP's Shapley-value foundation provides local accuracy, consistency, and missingness guarantees that are attractive for adverse-action reasoning and validation. Recent finance-specific work argues that explainability must be treated as a first-class, logged system property rather than an offline visualization [28], [29], a position this paper operationalizes.

2.3 Fairness

Fair-lending obligations require evaluation for disparate impact and group-level error disparities. The fairness literature formalizes competing criteria (demographic parity, equalized odds, and equal opportunity) and establishes that they cannot in general be satisfied simultaneously with calibration [10], [12], [13], [26], [27]. Interventions are commonly grouped into pre-processing (reweighting or transforming data), in-processing (fairness-constrained training), and post-processing (group-specific decision thresholds). We adopt a post-processing approach because it controls outcomes without altering an independently validated model, which is preferable under model-risk governance and easier to audit.

2.4 Responsible AI and Socio-Technical Perspectives

A growing body of work argues that fairness and accountability cannot be reduced to a model property and must be analyzed at the level of the socio-technical system in which a model is embedded: its data, institutions, incentives, and affected people [41]. Catalogues of AI-ethics guidelines converge on a small set of principles (beneficence, non-maleficence, autonomy, justice, and explicability) [44], [45], and finance-specific studies show that automated credit models can widen or narrow disparities depending on design and oversight [42], [43]. Parallel debates concern whether high-stakes decisions should rely on inherently interpretable models [46] or on post-hoc explanations [6], [48], and on documentation practices such as model cards [47]. This paper operationalizes these perspectives for CLI rather than treating them as external commentary.

2.5 Calibration and Model Governance

Threshold-based risk control is only as reliable as the PD estimates it consumes; raw classifier scores, especially under class-imbalance handling, are often mis calibrated [22], [23]. Platt scaling [20] and isotonic regression [21] are standard remedies. At the systems level, work on technical debt and production readiness in ML [34], [35], [36], [37] and on concept-drift adaptation [38] motivates the monitoring and registry components of our architecture, while SR 11 – 7 [14] and Basel guidance [15], [16] define the governance and capital context.

2.6 Research Gap

Although credit-scoring accuracy, explainability, and fairness are each well studied, little work frames proactive CLI as a constrained economic optimization, couples calibrated PD estimation to an explicit loss budget, controls fairness at the threshold with a measured economic cost, and embeds the result in a governed, reproducible socio-technical system. This paper targets that intersection.

3 Problem formulation

3.1 Account-Level Incremental Profit

For account i with feature vector x_i , let Δi denote the proposed credit-line increase, PD_i the probability of default over horizon T , LGD_i the loss given default, γ the net margin rate, and U_i the expected incremental utilization fraction. CLI economics are marginal: only the exposure created by the increase enters the trade-off. The incremental exposure, revenue, and expected loss are

$$EAD_i^{n\Delta} = \Delta i U_i, Ri = \gamma \Delta i U_i, Li = PD_i \cdot LGD_i \cdot EAD_i^{n\Delta} \quad (1)$$

so the expected incremental profit is $\Pi_i = Ri - Li = EAD_i^{n\Delta}(\gamma - PD_i LGD_i)$. An account is intrinsically profitable to expand whenever $PD_i < \gamma / LGD_i$, which defines a natural risk break-even point.

3.2 Portfolio Constraint and Objective

Institutions cap aggregate expected loss to reflect capital and risk-appetite limits. Over the eligible set S , the program solves

$$\max \sum_{i \in S} \Pi_i \text{ s.t. } \sum_{i \in S} PD_i LGD_i EAD_i^{n\Delta} \leq EL_{target} \quad (2)$$

together with fairness tolerances (Section 6) and operational caps. Continuous per-account optimization of Δi is expensive at portfolio scale; we therefore standardize the offered increment and reduce (2) to a calibrated PD threshold

$$Approve_i = \mathbf{1}[PD_i < \tau], \tau \text{ chosen so } \sum_{i: PD_i < \tau} PD_i LGD_i EAD_i^{n\Delta} \leq EL_{target} \quad (3)$$

This transforms the program into a calibrated classification problem under an explicit loss budget: prediction supplies the ranking, calibration makes the budget meaningful, and the threshold enforces it. The standardized increment also isolates decision quality from exposure heterogeneity when comparing models.

4 Data architecture and feature engineering

4.1 System Context and Synthetic Portfolio

The framework assumes a lakehouse environment (e.g., Delta tables) supporting versioned snapshots, reproducible feature pipelines, and event-time processing across four domains: transactions (merchant

category, timestamp, amount, channel), payments (statement balance, minimum due, amount paid, timing), bureau signals (score, score delta, inquiries, external utilization), and account master data (limit, tenure, product, revolving status).

Because account-level credit data cannot be shared, and to make every claim in this paper independently reproducible, we evaluate on a documented synthetic portfolio of $N = 60,000$ accounts rather than on proprietary data. A latent financial-stress factor and an unobserved idiosyncratic factor drive both behavior and default; the unobserved factor caps achievable discrimination at realistic levels. Nineteen features are generated as noisy proxies of stress across the four families below, and a protected-attribute proxy A is correlated with limit and score to create a realistic fairness challenge. Default is generated from a nonlinear risk index dominated by threshold and interaction terms (e.g., high utilization combined with weak payment), so that nonlinear models have structure to exploit beyond a linear model. The resulting base default rate is 18.1% over the label horizon (16.5% and 20.5% across the two protected groups). The generator and all code are released (Section 13).

4.2 Event-Time Features and Leakage Control

Transaction systems are subject to late payments, reversals, backfills, and bureau-refresh delays, so features are constructed on event-time windows ($t-3m$, $t-6m$, $t-12m$) rather than ingestion time to avoid training-serving skew. Labels are forward-looking over a 6–12 month horizon, $X_i(t) \rightarrow Y_i(t + T)$, with default defined as 60+ days past due or charge-off. Leakage is controlled by a strict forward label window, exclusion of post-decision behavior, time-based cross-validation, and bureau-lag adjustment.

4.3 Feature Families

Behavioral risk is driven by trend, stability, and drift rather than single-point levels. *Utilization dynamics* capture rolling mean utilization, the utilization slope (trend), and utilization volatility; a rising slope signals liquidity stress. *Payment stability* includes the payment-to-balance ratio, consecutive on-time payments, the full-to-minimum payment ratio, and payment-timing deviation, capturing repayment discipline beyond delinquency status. *Behavioral stability* includes spend coefficient of variation, revolver consistency, and monthly spend entropy. *Risk-drift indicators* include 3- and 6-month credit-score deltas, recent inquiries, and a short-term delinquency-spike flag. All feature definitions, time windows, and snapshot timestamps are version-controlled in the feature store so that training datasets record the exact feature version and data cutoff.

5 Modeling and calibration

The logistic baseline models log-odds linearly, $\log[p/(1 - p)] = \beta_0 + \sum_j \beta_j x_j$. The gradient-boosting model minimizes binary log-loss

$$L = - \sum_i [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] \quad (4)$$

where $\hat{p}_i$ is the predicted probability, fitting shallow trees (depth 5, learning rate 0.05, 600 rounds with early stopping, L2 regularization, subsampling) on the training fold and selecting rounds on the validation fold. Class imbalance is handled with positive-class weighting, which improves ranking but inflates raw probabilities. Because the loss budget in (3) requires PD estimates that match observed default frequencies, we calibrate the raw scores with isotonic regression fitted on the validation fold, enforcing $P(Y = 1) \approx$ empirical default rate. Model stability is tracked through the Population Stability Index (PSI), temporal AUC degradation, and feature-importance shift.

6 Explainability and fairness

6.1 SHAP-Based Explanation

For a prediction $f(x)$, SHAP decomposes the output additively as $f(x) = \phi_0 + \sum_j \phi_j$, where ϕ_0 is the baseline expectation and each ϕ_j is a feature's Shapley contribution [6], [7]. This yields local accuracy, consistency, and missingness. For every decision the system logs the model version, feature snapshot, calibrated PD, applied threshold, the SHAP vector, and the outcome, and translates the top contributors into human-readable reason codes (e.g., "consistent on-time payments reduced risk," "rising utilization increased risk") for audit, regulatory response, and customer disclosure. Mean absolute SHAP values aggregate to global drivers used for policy and strategy.

6.2 Fair-Lending Context and Fairness Criteria

Fair-lending law distinguishes disparate treatment (explicit use of a protected attribute) from disparate impact (a facially neutral policy with unjustified adverse effect on a protected group) [11]. Even when protected attributes are excluded from the model, correlated features can act as proxies, so disparate impact must be tested empirically. A common screening heuristic is the four-fifths rule, under which a disparate-impact ratio below 0.8 warrants scrutiny. The Fair Credit Reporting Act further requires that adverse decisions be accompanied by specific principal reasons, linking fairness to explainability.

Let A be a protected-attribute proxy. We monitor three criteria: the demographic-parity difference $DP = |P(\text{approve}|A = 0) - P(\text{approve}|A = 1)|$; equalized odds, comparing approval rates separately among truly non-defaulting accounts (an equal-opportunity / true-positive view) and among defaulting accounts (a false-positive view); and the disparate-impact ratio $DI = P(\text{approve}|A = 1)/P(\text{approve}|A = 0)$. These criteria can conflict, and none is universally correct, so the choice is a documented policy decision rather than a purely technical one.

6.3 Threshold Recalibration as a Governed Control

Rather than altering model structure, we recalibrate group thresholds to equalize approval rates subject to the loss budget, solving $\min_{\tau} |EL(\tau) - EL_{\text{target}}|$ s.t. $|EO| \leq \epsilon$. This post-processing choice has three governance advantages: it leaves the independently validated PD model unchanged, it exposes the economic cost of each fairness target explicitly (Section 7.5), and it is transparent to validators and

regulators. Because the intervention operates on the decision boundary, the institution can trace and justify exactly how much, and for whom, access was adjusted.

7 Experimental results

7.1 Protocol

We use a strict time-aware split: we train on years 1–2 (30,121 accounts), validate on year 3 (14,974), and test on year 4 (14,905), so that all reported metrics are out-of-time. The logistic baseline and gradient-boosting model are trained identically; calibration and thresholds are fit on validation and frozen before scoring the test fold. All numbers below are produced by the released code under a fixed random seed.

7.2 Discrimination and Calibration

Table 1 reports out-of-time discrimination. Gradient boosting improves AUC only from 0.758 to 0.762 and the Kolmogorov–Smirnov statistic from 0.404 to 0.419. Despite the threshold and interaction terms in the data-generating process, the regularized linear model recovers almost as much signal, because many noisy proxies of the same latent stress factor carry largely additive information and a strong unobserved factor bounds achievable discrimination; Fig. 2 shows the corresponding ROC curves.

Table 1 OUT-OF-TIME DISCRIMINATION (TEST FOLD)		
Model	AUC	KS
Logistic regression (baseline)	0.758	0.404
Gradient boosting + isotonic	0.762	0.419

Calibration is essential because the loss budget in (3) is enforced on PD magnitudes, not rankings. Table 2 shows that imbalance-weighted raw scores are badly miscalibrated (Brier 0.165, ECE 0.194); isotonic regression restores agreement between predicted and observed default rates (Brier 0.119, ECE 0.006). Figure 3 shows the reliability diagram before and after calibration.

Table 2 CALIBRATION QUALITY (TEST FOLD)		
Probabilities	Brier score	ECE
Raw (uncalibrated)	0.165	0.194
Isotonic-calibrated	0.119	0.006

7.3 Global Drivers and Ablation

Global mean-absolute-SHAP attribution (Fig. 4) ranks utilization slope, three-month utilization, payment-to-balance ratio, six-month score delta, and consecutive on-time payments as the dominant risk drivers, which are behaviorally sensible signals. The leave-one-family-out ablation in Table 3 confirms this: removing utilization features causes by far the largest AUC degradation, payment and risk-drift features contribute moderately, and behavioral-stability features add nothing once the others are present.

Table 3
ABLATION: AUC DROP WHEN A
FEATURE FAMILY IS REMOVED

Removed family	AUC drop
Payment stability	0.0106
Utilization dynamics	0.0419
Risk drift	0.0111
Behavioral stability	-0.0019

7.4 Economic Outcome

We set EL_{target} to 60% of the loss that approving every intrinsically profitable account ($PD < \gamma/LGD = 0.225$) would incur, yielding a calibrated threshold $\tau^* = 0.100$ and a 51.2% approval rate. Holding the approval rate fixed, we compare the gradient-boosting policy against the logistic policy on realized test outcomes (actual defaults), using a standardized increment so that the comparison reflects decision quality rather than exposure size. The two policies perform almost identically: the boosting policy's approved default rate is 8.56% against 8.46% for the logistic policy, its realized loss is 1.2% higher, and its net incremental profit is 1.1% lower (Table 4, Fig. 5). At this operating point a small AUC difference does not translate into an economic advantage.

Table 4
REALIZED PORTFOLIO ECONOMICS AT MATCHED APPROVAL (51.2%)

Policy	Approved default rate	Realized loss	Net profit
Logistic threshold	8.46%	258,000	370,290
Gradient boosting	8.56%	261,200	366,370
Relative change	+ 0.10 pp	+ 1.2%	-1.1%

7.5 Fairness, Group Performance, and the Cost of Parity

A central socio-technical question is whether disparities arise from the model performing worse for some group or from how a uniform decision rule allocates outcomes. Table 5 separates these: the PD model is essentially equally accurate and equally well-calibrated across the two protected groups (AUC 0.749 vs. 0.775; ECE 0.008 vs. 0.005), so the disparity is allocative, not a model-quality defect. Under a

single threshold the disadvantaged group (A1), which has a higher base default rate, receives a markedly lower approval rate (46.1% vs. 54.5%).

Table 5
PER-GROUP MODEL PERFORMANCE AND OUTCOMES (SINGLE THRESHOLD)

Group	AUC	ECE	Approval	Appr. default
A0 (advantaged)	0.749	0.008	54.5%	8.34%
A1 (disadvantaged)	0.775	0.005	46.1%	8.96%

Table 6 reports fairness before and after group-wise threshold recalibration. The single-threshold policy exhibits disparate impact (DP difference 0.084, equal-opportunity difference 0.071, DI ratio 0.85, above the 0.8 screening heuristic but a clear gap). Recalibration under the same loss budget reduces the demographic-parity difference to 0.013, the equal-opportunity (true-positive) gap to 0.037, and the false-positive gap from 0.069 to 0.015, while realized net profit changes by only – 0.6% (364,210 against 366,370 simulation units).

Table 6
FAIRNESS BEFORE AND AFTER THRESHOLD RECALIBRATION

Metric	Before	After
Demographic-parity difference	0.084	0.013
Equal-opportunity (TPR) gap	0.071	0.037
False-positive-rate gap	0.069	0.015
Disparate-impact ratio	0.85	1.03

Figure 6 traces the fairness–economics frontier obtained by progressively closing the parity gap through extending access to the disadvantaged group while holding the advantaged group fixed. In this portfolio the frontier slopes the favorable way: realized net profit rises from about 0.366 to 0.384 million simulation units as the demographic-parity gap closes from 0.084 to zero. Because this frontier adds approvals rather than reallocating them, it relaxes the loss budget; under a fixed budget (Table 6) the cost of parity is the 0.6% reported above. The single-threshold policy had been declining creditworthy A1 applicants whose incremental expected profit was positive; extending access to them improves both equity and economics. This is portfolio-specific: fairness can be costly when a disadvantaged group is genuinely riskier at the margin. Still, it demonstrates that the equity–profit relationship is an empirical question the framework makes measurable, not a foregone trade-off.

7.6 External Validation on a Real Public Benchmark

Because the main study uses synthetic data, we replicate the modeling and fairness pipeline on a real, publicly available credit dataset, the “credit_data” benchmark (4,454 applicants, 28.2% bad-credit rate) distributed via the Rdatasets archive [50], [51], using a stratified 60/20/20 split. The main pattern transfers (Table 7): gradient boosting and the logistic baseline discriminate equally (AUC 0.796 vs. 0.798; KS 0.472 vs. 0.466). With imbalance weighting the raw boosted probabilities are again miscalibrated (ECE 0.091), and isotonic calibration fitted on the small 891-applicant validation fold reduces this only to 0.063, which shows that reliable calibration needs adequate validation data. Treating applicant age, an ECOA-protected basis, as the protected attribute (under 30 vs. 30 and older), the single-threshold policy showed modest disparate impact at a 72.5% approval rate (demographic-parity gap 0.042, disparate-impact ratio 0.94), which group-wise recalibration reduced to 0.018 (ratio 1.03), although the false-positive gap widened, a reminder that fairness criteria conflict; the same diagnostic and control machinery therefore operates unchanged on real data.

Table 7
EXTERNAL VALIDATION ON REAL credit_data BENCHMARK
(TEST FOLD)

Metric	Logistic	Gradient boosting
AUC	0.798	0.796
KS statistic	0.466	0.472
ECE (raw → calibrated)	n/a	0.091 → 0.063
Disparate-impact ratio (age)	n/a	0.94 → 1.03

7.7 Multi-Period Feedback Dynamics

Credit decisions compound: an increase changes future utilization, repayment, and credit history, which feed back into later decisions. To examine these dynamics we run a stylized closed-cohort simulation of 20,000 accounts over eight periods in which approved increases raise limits and utilized balances, defaults realize and remove accounts, and borrower stress drifts in response to access (moderate new access reduces stress; over-extension increases it). We compare the single-threshold policy against the fairness-aware (group-equalized) policy. The dynamics make the socio-technical stakes visible (Fig. 7): under a single threshold the credit-access gap between groups widens across the horizon (from about 1,580 to 2,270 limit units, with the approval gap falling only from 0.13 to 0.10), whereas the fairness-aware policy holds the approval gap at zero and ends with a gap of about 1,300, 43% smaller. The long-run cost is small but non-zero: cumulative profit is about 0.6% lower under the fairness-aware policy. This tempers the static single-period frontier of Section 7.5 (where closing the gap was locally profit-positive) by showing that sustained, repeated extension of credit to a higher-risk group does carry a measurable multi-period cost, even though that cost remains small relative to the equity gained. The contrast underscores that fairness in lending must be evaluated dynamically, not only at a single decision point.

8 Production architecture and governance

The framework is deployed as a governed ML system (Fig. 1) comprising lakehouse ingestion, a versioned feature pipeline and store, the training module, a model registry with lineage, scoring and explainability services, a monitoring dashboard, and an immutable audit log. Drift is detected with the Population Stability Index, $PSI = \sum (p_i - q_i) \ln(p_i/q_i)$, with the usual bands ($PSI < 0.1$ stable; $0.1-0.25$ moderate; > 0.25 triggers retraining). The design maps directly onto SR 11 – 7 expectations [14]: documented conceptual soundness, independent validation enabled by logged feature snapshots and SHAP vectors, ongoing monitoring, and change-management records.

9 Ethical AI governance and socio-technical considerations

Because a CLI system distributes financial opportunity, its design choices are value choices. This section makes those choices explicit by mapping the framework to widely cited responsible-AI principles [44], [45], [49] and to the stakeholders a lending decision affects.

9.1 Stakeholders and Competing Values

Four stakeholders hold distinct, sometimes conflicting interests. The borrower seeks access to credit, fair treatment, and an understandable, contestable decision. The institution seeks profitable growth within risk appetite. The regulator seeks safety, soundness, and compliance with fair-lending and consumer-protection law. Society has an interest in equitable access to credit and in financial stability. The constrained-optimization formulation in Section 3 makes these tensions explicit: the objective encodes institutional profit, the loss budget encodes prudential and regulatory limits, and the fairness tolerance encodes borrower and societal equity. Rather than hiding these trade-offs inside a single score, the framework surfaces them as named, tunable parameters that a governance body can deliberate over and document.

9.2 Transparency and Contestability

Transparency operates at two levels. Globally, mean-absolute-SHAP attributions and the ablation analysis document what drives the model, supporting validation and policy review. Locally, every decision is accompanied by the calibrated PD, the applied threshold, and the SHAP contribution vector, which is translated into specific principal-reason codes. This supports the adverse-action explanations required under the Fair Credit Reporting Act and enables contestability: a borrower can be told why a request was declined and what would change the outcome, echoing the counterfactual-explanation literature [48]. Because explanations are logged with the model version and feature snapshot, they are reproducible after the fact rather than reconstructed.

9.3 Accountability and Human Oversight

Accountability requires that responsibility for automated decisions remain locatable. The model registry, immutable audit log, and versioned feature store record who deployed which model on what data and

which decision each borrower received, aligning with model-card-style documentation [47] and SR 11 – 7 governance [14]. The framework is designed for human oversight rather than full automation: governance committees set the loss budget and fairness tolerances, validators independently review conceptual soundness and monitoring, and material or borderline decisions can be routed for human review. Automation handles scale; humans retain authority over the value-laden parameters.

9.4 Feedback Loops and Distributional Dynamics

Credit decisions are not one-shot: an increase changes future utilization, repayment, and credit history, which in turn change later decisions. Such feedback loops can entrench advantage, repeatedly expanding low-risk customers while never giving marginal applicants the history that would qualify them, or, if designed responsibly, can expand access and build credit for underserved groups. Our multi-period simulation (Section 7.7) makes this concrete: under a single threshold the access gap persists across time, whereas a fairness-aware policy closes most of it at a modest long-run cost. The favorable static frontier of Section 7.5 thus understates the dynamic price of equity, which reinforces the need to monitor distributional effects over time rather than at a single decision point, and to test for proxy discrimination and intersectional disparities that a single binary attribute cannot capture.

9.5 Financial Inclusion

Taken together, the results support a measured claim: governed, fairness-aware machine learning can advance financial inclusion rather than only optimize loss. The same calibrated PD estimates that enforce a prudential loss budget also identify creditworthy applicants whom a coarse policy overlooks, and the threshold-level fairness control lets an institution expand access deliberately and auditably. Inclusion is thus treated not as charity that sacrifices soundness but as a design objective that the framework can pursue while respecting risk limits.

10 Capital and regulatory linkage

CLI directly affects regulatory capital. Under Basel internal-ratings-based approaches [15], [16], the capital requirement $K = f(\text{PD}, \text{LGD}, \text{EAD}, M)$ and expected loss $\text{EL} = \text{PD} \cdot \text{LGD} \cdot \text{EAD}$ both rise with the incremental exposure a CLI creates, while unexpected loss drives risk-weighted assets. Unconstrained expansion can therefore inflate RWA, compress capital ratios, and attract supervisory scrutiny. The loss budget in (3), $\sum \text{PD}_i \text{LGD}_i \text{EAD}_i^n \leq \text{EL}_{\text{target}}$, is a direct, auditable proxy for capital-planning discipline: institutions translate risk appetite (target charge-off rate, loss ceiling, capital buffer) into $\text{EL}_{\text{target}}$, and threshold calibration enforces it. The framework thus serves simultaneously as a revenue optimizer, a capital-control mechanism, and a documented bridge between AI deployment and prudential regulation.

11 Discussion

Five findings generalize beyond the specific numbers. First, model choice mattered far less than decision design: gradient boosting added little discrimination and no profit over a logistic baseline on either

dataset, while calibration and the threshold policy changed outcomes materially. Second, CLI is best modeled as constrained allocation, not classification: a moderate-PD account can be profitable when incremental revenue exceeds incremental expected loss, so the break-even PD, not an arbitrary cutoff, anchors the decision. Third, calibration is not cosmetic; without it the loss budget is enforced on the wrong scale, and the large calibration error we observe under imbalance handling shows how easily raw scores violate the constraint. Fourth, fairness in this setting is primarily an allocative choice at the decision boundary rather than a model-accuracy problem: the PD model was equally accurate across groups, yet a uniform threshold produced disparate impact, and controlling it at the threshold makes the intervention transparent and its cost measurable. Fifth, equity and profitability are not inherently opposed: our frontier shows a region where extending access to creditworthy members of the disadvantaged group improved both, so the trade-off is an empirical quantity to be measured per portfolio, not assumed. The system-level contribution (feature governance, training–serving parity, drift detection, explainability, fairness control, economic simulation, and capital linkage in one pipeline) distinguishes this work from accuracy-only credit studies and situates it within the responsible-AI and software-engineering-for-ML literatures [34]–[37], [41], [44].

12 Limitations and threats to validity

The principal limitation is that the main study uses synthetic data, so absolute metrics reflect the generator’s assumptions; the small logistic-versus-boosting gap, calibration behavior, and fairness–economics relationships are the transferable findings, not the specific values. We partially mitigate this with external validation on a real public benchmark (Section 7.6), where the near-parity of the two models and the fairness diagnostics reproduce, but that dataset describes loan-approval rather than credit-line increases and uses age as the protected attribute; validation on real revolving-credit data with richer protected attributes remains future work. The fairness analysis uses binary protected-attribute proxies and approval-rate parity; real deployments must address intersectional groups, proxy discrimination, the choice among incompatible fairness criteria as a documented policy decision, and jurisdiction-specific fair-lending tests. The multi-period analysis (Section 7.7) is a stylized closed-cohort simulation with an assumed behavioral-feedback model and no new-account inflow or macroeconomic dynamics; it illustrates qualitative distributional effects rather than calibrated long-run magnitudes. Finally, the economic model assumes fixed margin, LGD, and incremental-utilization parameters; sensitivity analysis and causal uplift modeling of behavioral response are left to future work.

13 Reproducibility

All results are generated by seed-fixed scripts (a single command, `python run_all.py`, regenerates every number) that produce the synthetic portfolio, train both models, calibrate probabilities, compute SHAP attributions, run the fairness and ablation analyses, execute the economic simulation, render every figure and table value in this paper, replicate the pipeline on the real public benchmark of Section 7.6, and run the multi-period feedback simulation of Section 7.7. The data generator, feature definitions, hyperparameters, calibration and threshold logic, external-validation and dynamics code, and plotting

code are released (see Data availability) so that every analysis can be reproduced end to end and extended to additional real portfolios.

14 Conclusion and future work

We presented a fairness-aware, explainable, and governed framework for proactive credit-line increases, treated as a responsible-AI problem in a socio-technical lending system. The framework formalizes CLI as constrained incremental-profit maximization under a portfolio loss budget and fairness tolerances, enforces the budget through calibrated PD thresholding, controls disparate impact with auditable group-level recalibration, attaches SHAP-based reason codes for transparency and contestability, and deploys within a monitored, accountable architecture mapped to responsible-AI principles and prudential governance. In a reproducible simulation study, gradient boosting matched but did not materially beat a logistic baseline, the PD model was equally accurate across groups, calibration sharply reduced probability error, threshold recalibration cut the demographic-parity gap from 0.084 to 0.013 at a 0.6% profit cost, and a fairness–economics frontier showed that extending access to creditworthy members of the disadvantaged group was profit-positive. Future work will pursue longitudinal evaluation of distributional feedback, intersectional and proxy-discrimination analysis, constrained reinforcement learning for continuous limit setting, causal uplift modeling of behavioral response, macroeconomic stress testing, and validation on public and proprietary portfolios.

Declarations

Competing interests.

The authors declare no competing interests.

Ethics approval and consent to participate.

Not applicable. The study uses synthetic data and a publicly available, de-identified benchmark; no human participants or personal data were involved.

Consent for publication.

Not applicable.

Funding.

No funding was received for this work.

Author Contribution

A.S. conceived the framework, developed the methodology and software, ran the experiments, and wrote the manuscript. G.S. contributed to the problem formulation and the fair-lending and governance analysis, and reviewed and edited the manuscript. Both authors read and approved the final manuscript.

Data Availability

The synthetic portfolio generator, feature definitions, analysis code, and all scripts needed to reproduce the results are available at https://github.com/akshaysharma1088/Credit_Line_Determination (and archived at Zenodo - 10.5281/zenodo.23710023). The public credit benchmark used in Section 7.6 is available through the Rdatasets archive [51].

References

1. Thomas LC, Crook JN, Edelman DB. Credit Scoring and Its Applications. 2nd ed. Philadelphia, PA, USA: SIAM; 2017.
2. Hand DJ, Henley WE. Statistical classification methods in consumer credit scoring: A review. *J Roy Statist Soc A*. 1997;160(3):523–41.
3. Lessmann S, Baesens B, Seow H-V, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur J Oper Res*. 2015;247(1):124–36.
4. Chen T, Guestrin C. XGBoost: A scalable tree boosting system, in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.
5. Ke G et al. LightGBM: A highly efficient gradient boosting decision tree. in *Adv Neural Inf Process Syst*, 30, 2017.
6. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–74.
7. Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.
8. Ribeiro MT, Singh S, Guestrin C. ‘Why should I trust you?’ Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD*, 2016, pp. 1135–1144.
9. Molnar C. *Interpretable Machine Learning*, 2nd ed. 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
10. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. in *Adv Neural Inf Process Syst*, 29, 2016.
11. Barocas S, Selbst AD. Big data’s disparate impact. *Calif Law Rev*. 2016;104(3):671–732.
12. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores, in *Proc. ITCS*, 2017.
13. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness, in *Proc. 3rd ITCS*, 2012, pp. 214–226.

14. Board of Governors of the Federal Reserve System and OCC. Supervisory guidance on model risk management. SR Letter 11 – 7; 2011.
15. Basel Committee on Banking Supervision. International Convergence of Capital Measurement and Capital Standards (Basel II). Basel, Switzerland: BIS; 2006.
16. Basel Committee on Banking Supervision. III: Finalising Post-Crisis Reforms. Basel: Basel, Switzerland: BIS; 2017.
17. Siddiqi N. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Hoboken, NJ, USA: Wiley; 2006.
18. Baesens B, et al. Benchmarking state-of-the-art classification algorithms for credit scoring. *J Oper Res Soc.* 2003;54(6):627–35.
19. Brown I, Mues C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst Appl.* 2012;39(3):3446–53.
20. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classifiers*, 1999, pp. 61–74.
21. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates, in *Proc. 8th ACM SIGKDD*, 2002, pp. 694–699.
22. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning, in *Proc. 22nd ICML*, 2005, pp. 625–632.
23. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks, in *Proc. 34th ICML*, 2017, pp. 1321–1330.
24. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Statist.* 2001;29(5):1189–232.
25. Yeh I-C, Lien C-H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst Appl.* 2009;36(2):2473–80.
26. Verma S, Rubin J. Fairness definitions explained, in *Proc. IEEE/ACM Int. Workshop Softw. Fairness (FairWare)*, 2018, pp. 1–7.
27. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv.* 2021;54(6):1–35.
28. Bücken M, Szepannek G, Gosiewska A, Biecek P. Transparency, auditability, and explainability of machine learning models in credit scoring. *J Oper Res Soc.* 2022;73(1):70–90.
29. Bracke P, Datta A, Jung C, Sen S. Machine learning explainability in finance, Bank of England, Staff Working Paper 816, 2019.
30. Gunnarsson V, vanden Broucke S, Baesens B, Óskarsdóttir M, Lemahieu W. Deep learning for credit scoring: Do or don't? *Eur J Oper Res.* 2021;295(1):292–305.
31. Crook JN, Edelman DB, Thomas LC. Recent developments in consumer credit risk assessment. *Eur J Oper Res.* 2007;183(3):1447–65.
32. Anderson R. The Credit Scoring Toolkit. Oxford, U.K.: Oxford Univ. Press; 2007.

33. Thomas LC. A survey of credit and behavioural scoring. *Int J Forecast*. 2000;16(2):149–72.
34. Sculley D et al. Hidden technical debt in machine learning systems. *Adv Neural Inf Process Syst*, 28, 2015.
35. Breck E, Cai S, Nielsen E, Salib M, Sculley D. The ML test score: A rubric for ML production readiness, in *Proc. IEEE Big Data*, 2017, pp. 1123–1132.
36. Amershi S et al. Software engineering for machine learning: A case study, in *Proc. IEEE/ACM ICSE-SEIP*, 2019, pp. 291–300.
37. Paleyes A, Urma R-G, Lawrence ND. Challenges in deploying machine learning: A survey of case studies. *ACM Comput Surv*. 2022;55(6):1–29.
38. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. *ACM Comput Surv*. 2014;46(4):1–37.
39. Hurley M, Adebayo J. Credit scoring in the era of big data. *Yale J Law Technol*. 2016;18:148–216.
40. Kvamme H, Sellereite N, Aas K, Sjörsen S. Predicting mortgage default using convolutional neural networks. *Expert Syst Appl*. 2018;102:207–17.
41. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. Fairness and abstraction in sociotechnical systems, in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT)*, 2019, pp. 59–68.
42. Bartlett R, Morse A, Stanton R, Wallace N. Consumer-lending discrimination in the FinTech era. *J Financ Econ*. 2022;143(1):30–56.
43. Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A. Predictably unequal? The effects of machine learning on credit markets. *J Finance*. 2022;77(1):5–47.
44. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389–99.
45. Floridi L, Cowls J. A unified framework of five principles for AI in society. *Harv Data Sci Rev*, 1, 1, 2019.
46. Rudin C. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206–15.
47. Mitchell M et al. Model cards for model reporting, in *Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT)*, 2019, pp. 220–229.
48. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv J Law Technol*. 2018;31(2):841–87.
49. Kearns M, Roth A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford, U.K.: Oxford Univ. Press; 2019.
50. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York, NY, USA: Springer; 2013.
51. Arel-Bundock V. *Rdatasets: A collection of datasets originally distributed with R packages*, 2024. [Online]. Available: <https://vincentarelbundock.github.io/Rdatasets/>

Figures

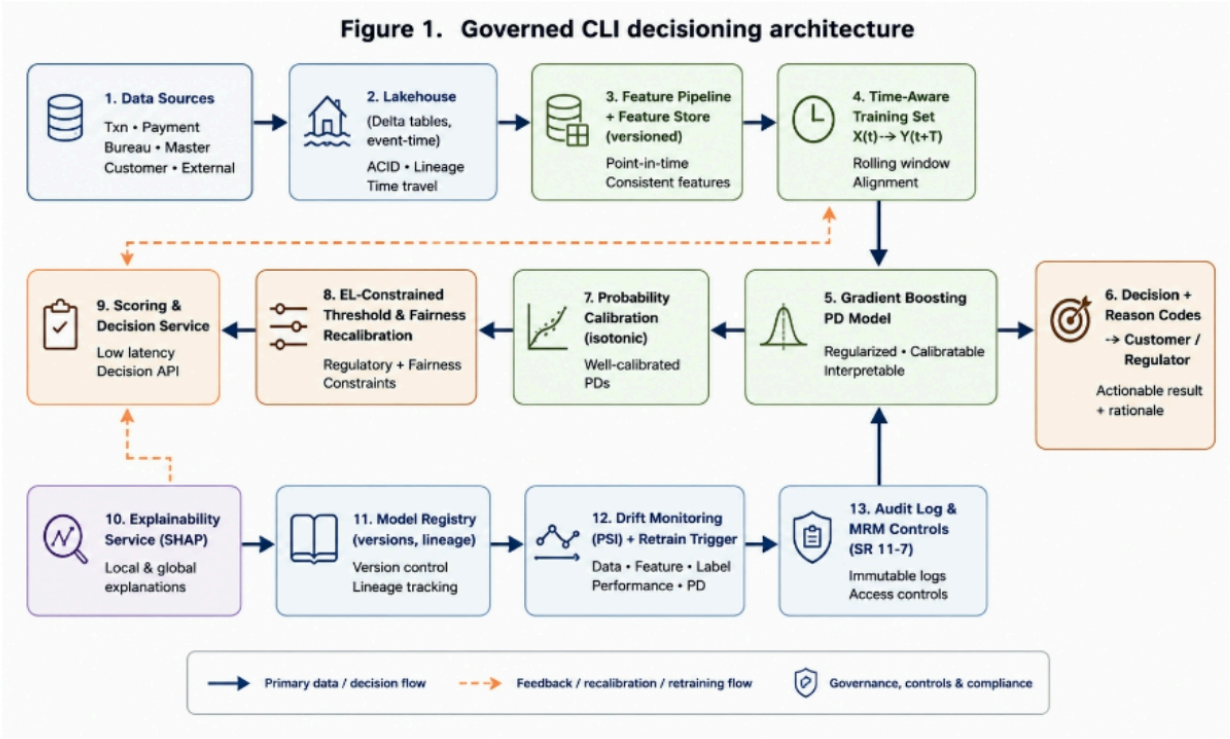


Figure 1

Governed CLI decisioning architecture: lakehouse ingestion and a versioned feature store feed a calibrated PD model; an expected-loss-constrained, fairness-aware threshold produces decisions and SHAP reason codes, with a registry, drift monitoring, and audit logging closing the governance loop.

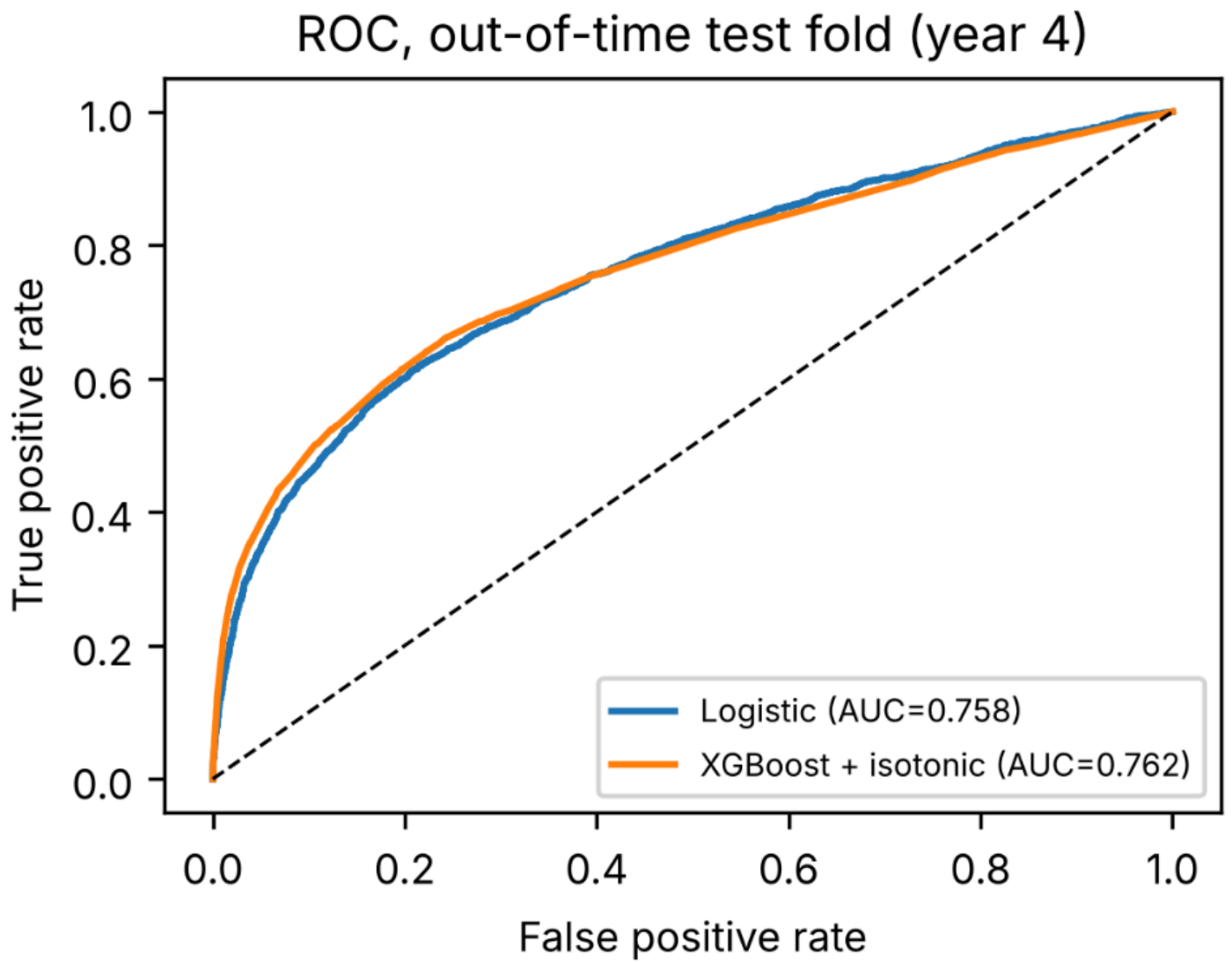


Figure 2

ROC curves on the out-of-time test fold.

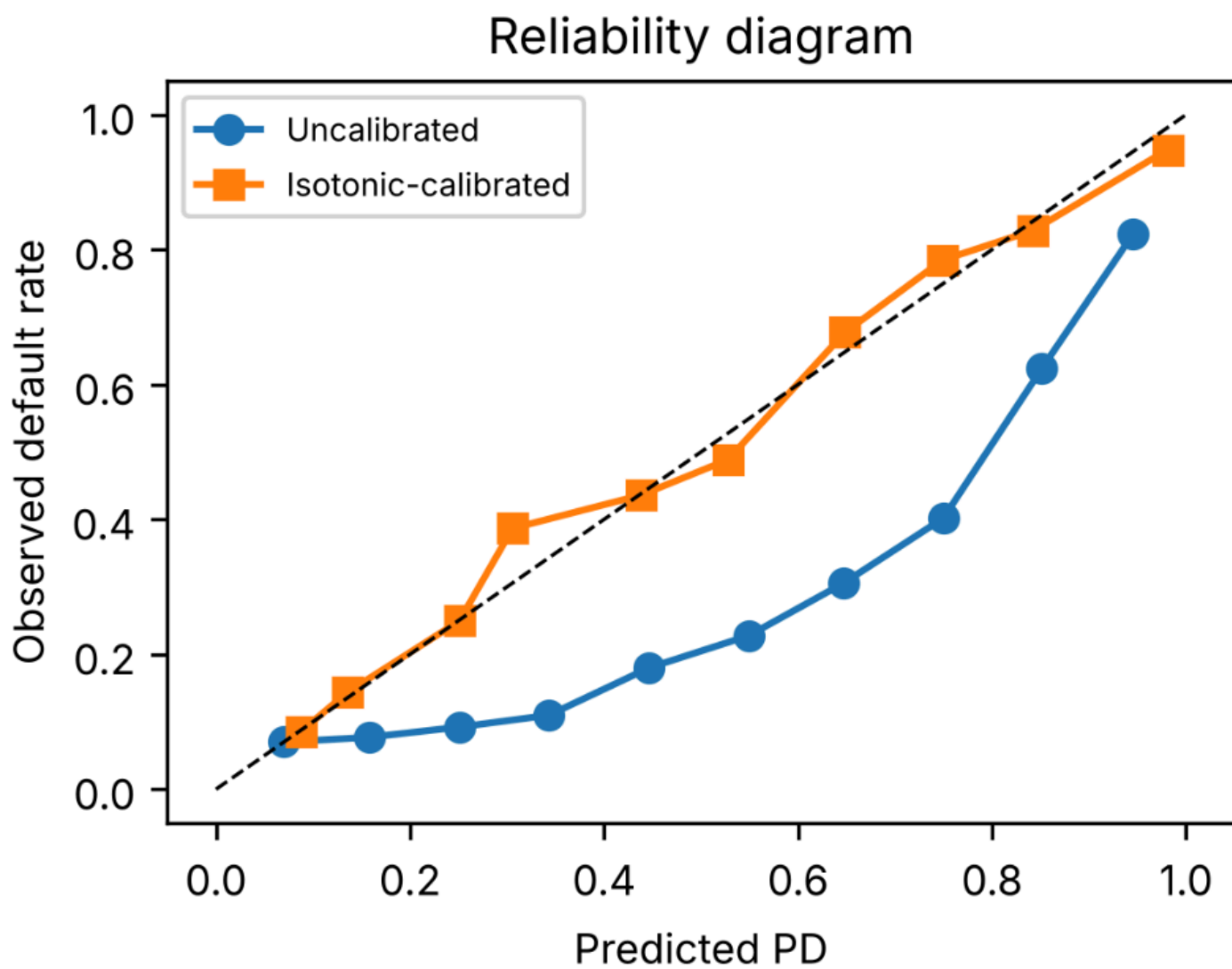


Figure 3

Reliability diagram before and after isotonic calibration.

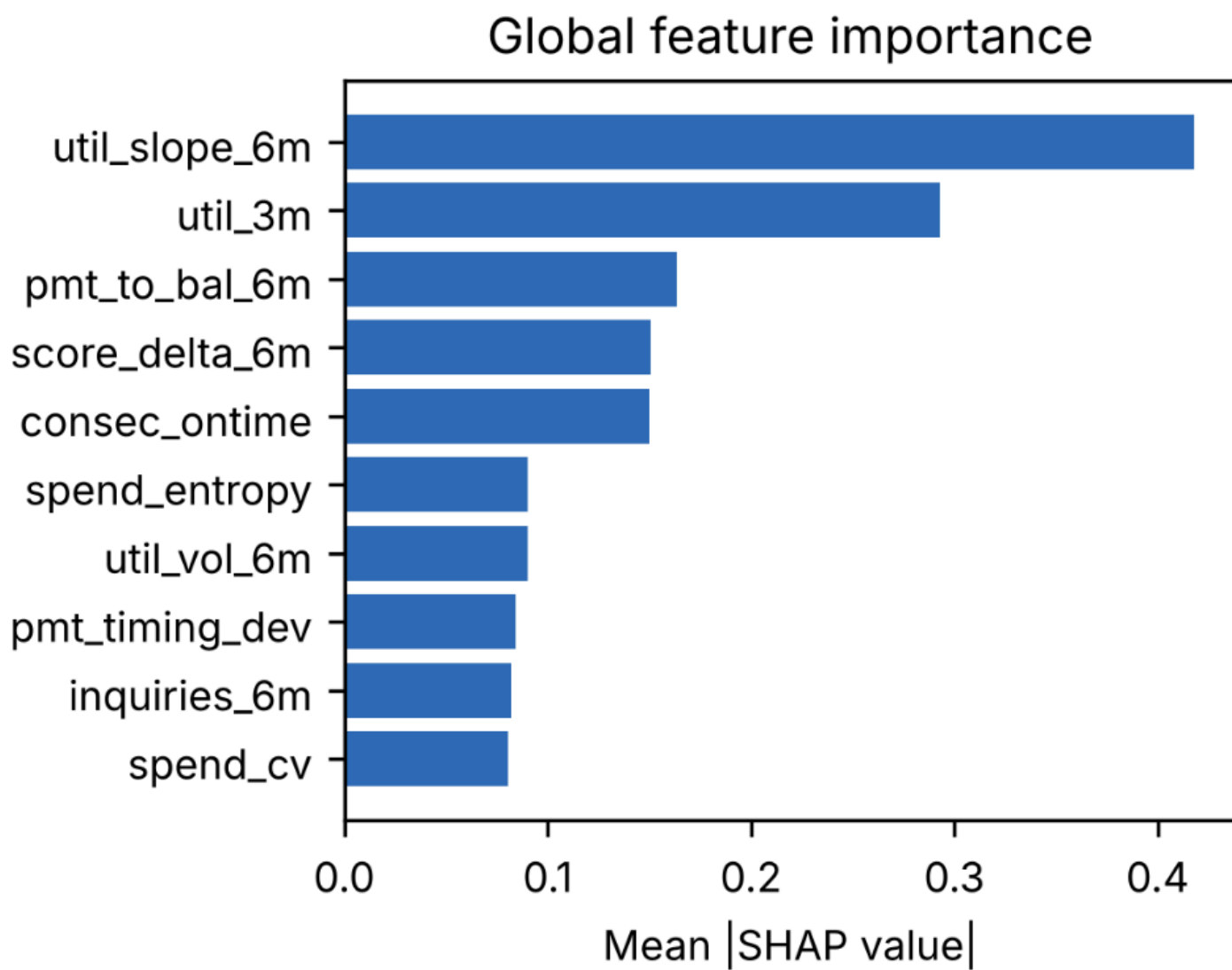


Figure 4

Global feature importance (mean |SHAP value|).

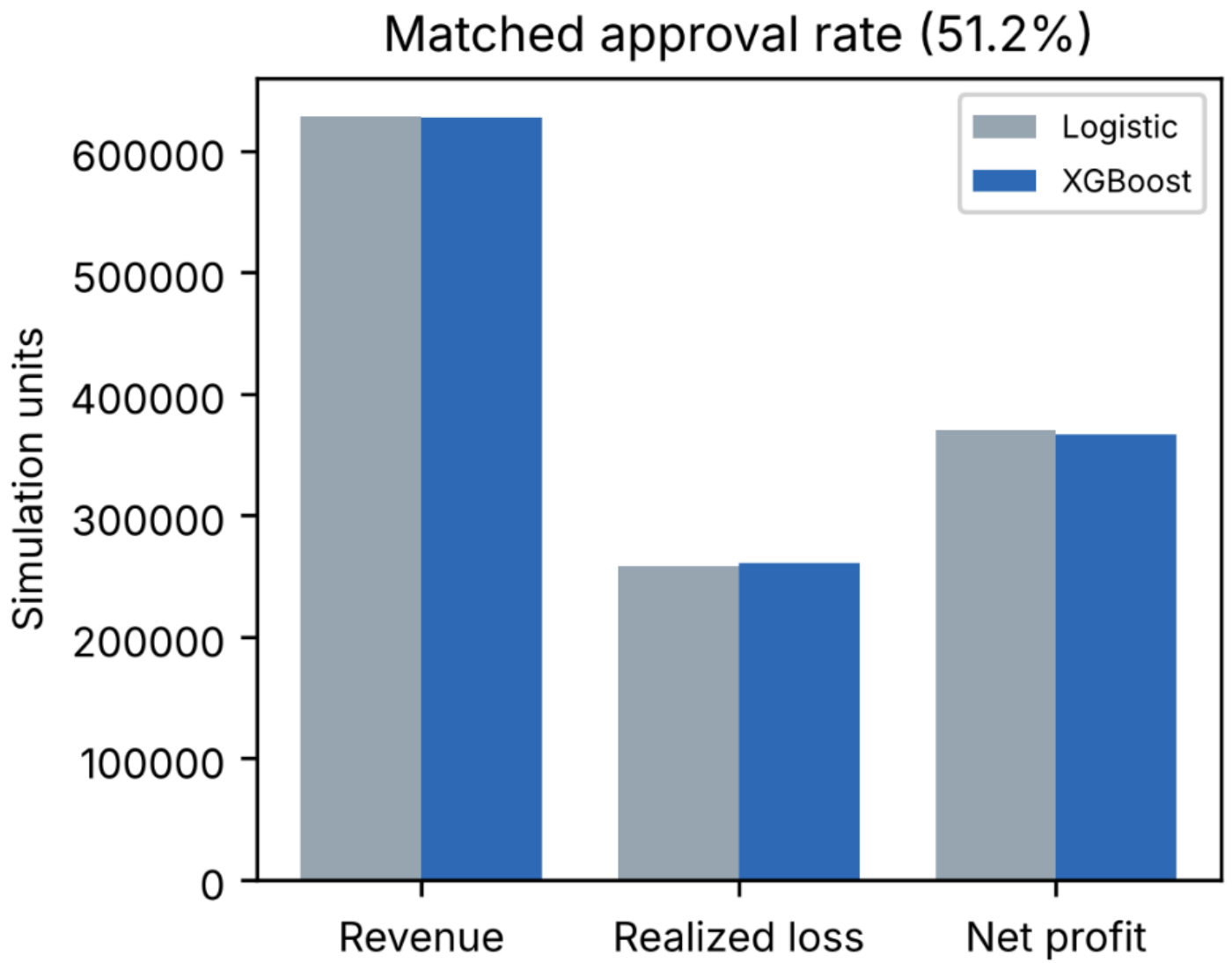


Figure 5

Realized revenue, realized loss, and net profit at the matched operating point (simulation units).

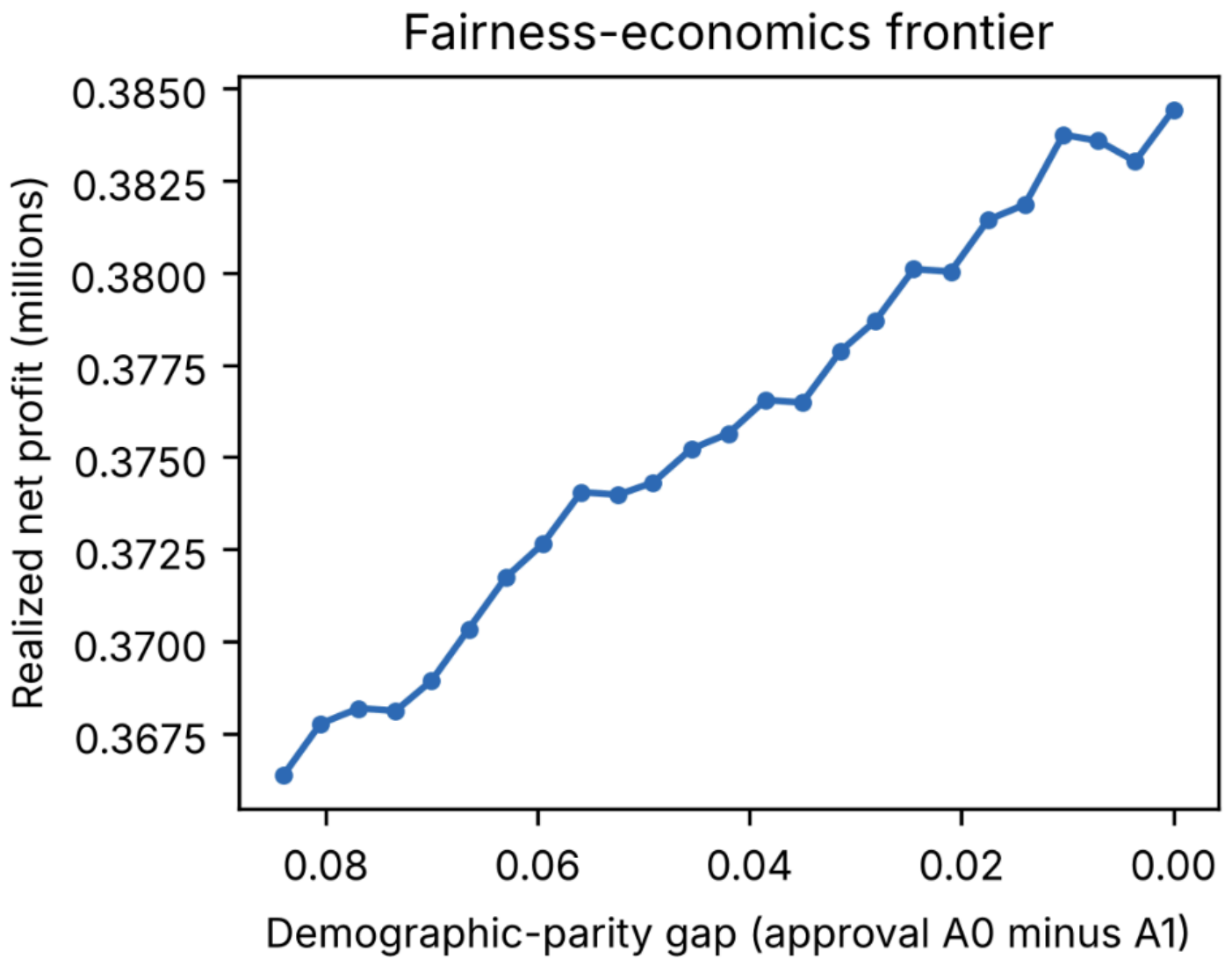


Figure 6

Fairness-economics frontier: realized net profit as the demographic-parity gap is closed by extending access to the disadvantaged group.

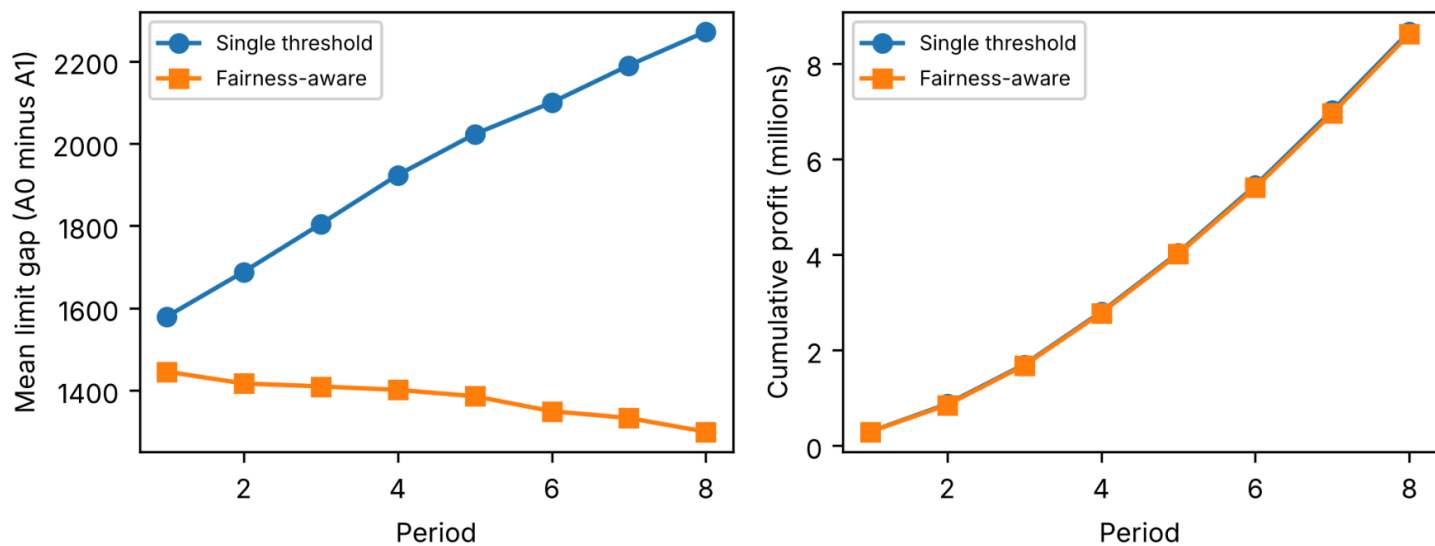


Figure 7

Multi-period dynamics under single-threshold vs. fairness-aware policies: (left) credit-access gap between groups over time; (right) cumulative profit over time.