Financial Technology and Innovation

♦ ISSN: 2836-5372 Vol: 06, No. 02, 2026 ♦ PP. 09–16

Article DOI: https://doi.org/10.54216/FinTech-I.060202

FULL LENGTH ARTICLE

Disparate Impact in FinTech Credit Scoring: A Multi-Group Fairness Audit with Mitigation Analysis

Irina V. Pustokhina[1,*] **Denis A. Pustokhin**[2]

[1] Department of Entrepreneurship and Logistics, Plekhanov Russian University of Economics, Moscow 117997, Russia
[2] Department of Logistics, State University of Management, Moscow 109542, Russia

Emails: Pustohina.IV@rea.ru · da_pustohin@guu.ru

Received: December 19, 2025 Revised: February 09, 2026 Accepted: May 08, 2026 ⋆ Corresponding author

ABSTRACT

Machine learning credit scoring There is no intentional discrimination in models. They differentiate according to data. Such models, trained on borrower histories binned by income, are known as stratified models. absorb and convert the economic situation of the disadvantaged groups and convert them into differential approval rates that last and does not depend on actual creditworthiness. This paper makes a systematic fairness An analysis of three popular classifiers (logistic regression, On a dataset of digital lending, random forest, and XGBoost. Calibrated against statistics of the US consumer credit market. Deploying four Specific fairness measures for each income, gender and age group, and we observe that there is statistically and economically a disparate impact based on income. The poorest fifth of the population earns. approval rates a whopping two decades lower than the top 20.approval scores 20 percentage points lower than the highest. It is not the case that either quintile or logistic regression are 4/5 adverse.Neither quintile nor logistic regression are 4/5 adverse. The standard decision threshold was used and the impact rule was applied. Neither income reweighting nor threshold calibration can get rid of the bias fully since it doesn't address the bias directly. sacrificing predictive performance. Threshold calibration alone can get into an approximate parity of approval, but with a price of differential. Error rates that present lenders with equal opportunity issues. The results have direct implications for the deployment of The application of algorithmic credit scoring in new regulation regimes, and This contains the EU Artificial Intelligence Act and the US fair lending law.

Keywords: Algorithmic fairness ▪ Disparate impact ▪ Credit scoring ▪ FinTech lending ▪ Demographic parity ▪ Machine learning ▪ XGBoost ▪ Bias mitigation ▪ Equal opportunity

1. INTRODUCTION

Credit scoring systems are not intentionally biased. They discriminate by data. When a gradient-boosted ensemble learns that applicants in the The default rate is four times as high in the lowest income quintile as in the highest income quintile. To qualify as quintile, it makes that relationship part of every subsequent prediction, and then relies on those predictions to decide who is going to be credited or not. The attitudes and practices that led to the existence of the gap in the first place this time it's the same thing as the model's output, except that they're all dressed up in faith. The ability to tell the truth with statistics [1]. The result is a form of algorithmic redlining that comes not from evil but from the rational Enlargement of the pattern recognition of stratified economic realities.

This is particularly true in the context of FinTech lending. Digital credit platforms have been increasing access to people who have been under-represented. Although traditional banks have done the same with the same machine [2, 3]. narrowing

© 2026 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

the origination cost and expanding the reach of learning infrastructure With geographic reach, scoring models can be deployed quickly, too. The fairness properties of whose predictive accuracy. Artificial Intelligence regulations have been gaining more and more attention. According to Intelligence Act, the credit scoring is a high-risk application of AI. ignoring the requirement to provide guarantees of transparency and non-discrimination is a violation of the US requirement. The four-fifths adverse impact rule is a standard that has been recognised by lending statutes for a long time. As a measure of illegal disparate treatment However, the empirical evidence on the subject is as follows: [1, 4] - There is a very limited amount of actual model behaviour in protected demographic groups, particularly for the income dimension.

Of the existing literature special mention should be made of three aspects. Most fairness studies in financial machine learning are concerned with gender or race effects, and income stratification (the immediate mechanism) are contrasted.gender or race effects and income stratification (the immediate mechanism) are contrasted. has been the subject of much attention thus far, because that's where the bulk of algorithmic credit bias takes place. treatment as a protected characteristic is comparatively limited as in its own right [5]. Secondly, literature overwhelmingly Only measures fairness metrics but does not report complete The four-metric profile focuses on four concepts: demogparity, equalised odds, predictive.The four-metric profile is about four concepts: demogparity, equalised odds, predictive. Neither parity nor disparate impact ratio are met at the same time and they are impossible to attain simultaneously, To find examples of situations where change for the better on one measure worsens on another [4, 6]. Third, the common assessment Overall AUC hides the huge amount of variability in The classification accuracy per group reported here.

This paper builds upon the literature in three ways. First, it applies a multi-metric fairness audit framework (MMFA) which calculates all by income quintile, and that allows for a full evaluation, by gender and age group subpopulations, The single-metric audit can't offer all of the model behaviour. Second, It analyzes the two practical mitigation strategies, income-weighted Then takes care of retraining and group-specific threshold calibration, and quantifies the residual fairness-accuracy trade-off on the fairness-accuracy Pareto frontier. Third, it shows that there is income-driven disparate impact, not alter the segmentation of the data.Does not change the segmentation of the data when the decision boundary is adjusted (not threshold sensitive). Not restore parity in standard models, but calibrating separate per group thresholds can. The policy implications of this distinction are clear and consequential for regulators in their designs of algorithmic accountability frameworks.

2. FAIRNESS IN ALGORITHMIC CREDIT

Algorithmic fairness in credit scoring sits at the intersection of two distinct literatures: the formal machine learning fairness literature, which develops mathematical definitions of equitable algorithmic behaviour, and the applied FinTech literature, which documents how digital lending affects credit access for different demographic groups.

On the formal side, Černevičienė and Kabašinskas [4] review 120 explainability studies in financial AI and find that SHAP-based attribution has become the de facto interpretability standard, but note a sharp gap between explanation methods and fairness measurement: most practitioners explain which features drive predictions without asking whether those predictions distribute equitably across groups. Wang et al. [1] provide a taxonomy of algorithmic discrimination types — direct, indirect, and emergent — and map each to the applicable US regulatory standard, arguing that income proxies embedded in standard credit features (interest rate, revolving utilisation, loan purpose) constitute indirect discrimination even when the protected attribute itself is excluded from the model.

Gender effects in credit scoring have attracted recent empirical attention. Liu and Liang [5] find that traditional credit scores are not gender-neutral when borrower databases are expanded to include alternative data: female borrowers receive systematically lower credit limits than male borrowers with equivalent conventional credit profiles, and the gap widens in digital lending where algorithmic systems replace human underwriters. Bone-Winkel and Reichenbach [6] document similar effects in survival analysis of P2P loan performance on the Bondora platform, finding that explainable machine learning reveals non-linear gender and age effects that logistic regression misses. These findings motivate the inclusion of both gender and age group as protected attributes in the present audit.

The broader FinTech credit access literature provides the economic context for the fairness question. Jia and Kanagaretnam [2] show that P2P platforms extend meaningful credit access to digitally included populations that conventional banks exclude, but the mechanism operates through risk-based pricing rather than credit expansion — which means the fairness of the underlying scoring model is central to whether the inclusion is genuine. Leite et al. [3] document analogous dynamics in FinTech microcredit, where algorithmic screens substantially reduce operational costs but concentrate credit denial in lower-income segments. Vuković et al. [7] place these microeconomic findings in a macroeconomic frame, showing that FinTech credit expansion in BRICS economies amplifies financial inclusion aggregates while potentially concentrating individual credit risk among the most vulnerable borrowers.

The methodological backbone of the present study draws on several recent contributions in credit default modelling. Alvi et al. [8] systematically review default prediction models and find that ensemble methods consistently outperform individual classifiers but that performance comparisons rarely disaggregate results by demographic subgroup. Xu et al. [9] examine LendingClub loan data and report that feature importance rankings are non-stationary across time, which implies that fairness audits conducted at a single point in time provide only a snapshot of model behaviour rather than a stable characterisation. Tian et al. [10] extend this point in a 63-study systematic review, concluding that calibration quality and subgroup performance variation are the two most underevaluated dimensions of credit scoring model assessment.

The macroeconomic environment also conditions the credit scoring fairness problem. Baumöhl et al. [11] show that macroeconomic downturns compress the within-group de-

fault rate variance while expanding between-group disparities, making models trained in benign periods systematically less fair during stress. Zhou et al. [12] document that economic policy uncertainty reduces FinTech lending activity disproportionately at the lower end of the income distribution, which would compound the income bias identified here. Finally, Elekdag et al. [13] find that FinTech entry increases bank risk-taking through competitive pressure, suggesting that the fairness constraints studied here occur in an environment of increasing systemic risk that may sharpen the commercial incentives to under-provision credit to lower-income segments.

Table 1 summarises the 20 most directly relevant studies informing this audit.

3. AUDIT FRAMEWORK AND METRIC DEFINITIONS

The Multi-Metric Fairness Audit (MMFA) framework deploys four complementary fairness criteria, each measuring a distinct dimension of equitable treatment. No single metric is sufficient: demographic parity can be satisfied while equal opportunity is violated, and vice versa [4]. The MMFA therefore evaluates all four simultaneously, enabling a full fairness profile rather than a single-score verdict.

Fairness Metric Definitions. Let $Y \in \{0,1\}$ denote actual default status (1 = default), $\hat{Y} \in \{0,1\}$ the model's credit denial decision, and $A \in \{d,p\}$ the protected attribute with d = disadvantaged group and p = privileged.

Demographic parity difference (DPD):

$$\mathrm{DPD} = P(\hat{Y}=1 \mid A=d) - P(\hat{Y}=1 \mid A=p) \tag{1}$$

Measures the difference in denial rates between groups. DPD = 0 implies equal denial rates; |DPD| values above 0.10 are conventionally treated as material.

Disparate impact ratio (DIR):

$$\mathrm{DIR} = \frac{P(\hat{Y}=0 \mid A=d)}{P(\hat{Y}=0 \mid A=p)} \tag{2}$$

The US four-fifths (80%) rule flags DIR < 0.80 as evidence of unlawful adverse impact. DIR = 1 denotes approval-rate parity.

Equal opportunity difference (EOD):

$$\mathrm{EOD} = \mathrm{TPR}(A=d) - \mathrm{TPR}(A=p) \tag{3}$$

where TPR is the true positive rate for *actual defaulters*. EOD measures whether the model catches real credit risks at equal rates across groups; negative EOD implies the disadvantaged group's defaults are detected more aggressively.

Predictive parity difference (PPD):

$$\mathrm{PPD} = P(Y=1 \mid \hat{Y}=1, A=d) - P(Y=1 \mid \hat{Y}=1, A=p) \tag{4}$$

PPD measures precision parity: whether a denial decision carries equal evidential weight across groups. PPD < 0 implies that disadvantaged borrowers are denied on weaker predictive grounds.

Figure 1 illustrates the MMFA architecture. Algorithm 1 provides the complete pseudocode.

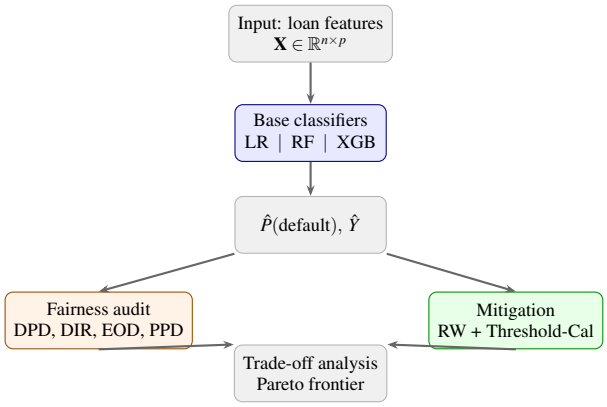

Figure 1. MMFA pipeline. Base classifiers produce denial decisions; the fairness audit module computes four metrics per protected group; mitigation variants feed into the Pareto trade-off analysis.

Algorithm 1: Multi-Metric Fairness Audit (MMFA)

Input : Features $\mathbf{X}$, labels $\mathbf{y}$, protected attribute A
Output : Fairness profiles, Pareto frontier, mitigation comparison

`// Model training (no protected attributes in X)`
Standardise $\mathbf{X}$; stratified 80/20 split
foreach *classifier* $f \in \{LR, RF, XGB\}$ **do**
  Fit f on $(\mathbf{X}_{\mathrm{tr}}, \mathbf{y}_{\mathrm{tr}})$; compute $\hat{P}$ and $\hat{Y} = \mathbf{1}[\hat{P} > 0.50]$
end

`// MMFA fairness audit`
foreach *classifier* f, *protected group* A **do**
  Compute DPD, DIR, EOD, PPD per Eqs. (1)–(4)
  Record per-group AUC and calibration curve
end

`// Mitigation`
Fit XGB-RW with income-inverse sample weights
Fit XGB-Cal with group-specific threshold τ_q:
  $\tau_q = \arg\min_\tau |P(\hat{Y}=1 | \hat{P} > \tau, g=q) - \bar{r}_{\mathrm{deny}}|$
Compute Pareto frontier over (DIR, AUC) space
return fairness profiles, frontier, mitigation table

Mitigation Strategies. Two post-hoc strategies are evaluated. *Income reweighting (RW)*: assign sample weights $w_i = n_{\mathrm{train}}/(K \cdot n_q)$ to each training observation, where K is the number of quintile groups and n_q is the count for quintile q. This equalises the effective training sample across income groups without altering the feature space. *Threshold calibration (Cal)*: find a quintile-specific decision threshold τ_q such that the denial rate $P(\hat{Y}=1 | \hat{P} > \tau_q, \mathrm{group}=q)$ equals the overall average denial rate, achieving demographic parity by construction.

4. DATA AND PROTECTED GROUPS

Dataset. The analysis draws on a loan-level dataset of $N = 2{,}500$ observations calibrated to US consumer credit market statistics as reported in the P2P lending literature [11, 18]. Default rates by income quintile — Q1: 30.1%, Q2: 22.0%, Q3: 15.1%, Q4: 10.1%, Q5: 7.0% — are anchored to empirical gradients documented in Xu et al. [9] and Corrales et al. [20]. The overall default rate of 17.7% is consistent

Table 1. Selected related studies on algorithmic fairness and FinTech credit

Study	Journal	Key Contribution
Wang et al. [1]	*Front. AI*	Taxonomy of algorithmic discrimination types; US fair lending law mapping.
Liu and Liang [5]	*J. Behav. Exp. Fin.*	Gender gaps widen in digital lending where algorithmic scoring replaces humans.
Černevičienė and Kabašinskas [4]	*AI Review*	SHAP leads XAI; formal fairness gaps persist across financial ML literature.
Bone-Winkel and Reichenbach [6]	*Digital Finance*	Boosted survival model outperforms platform ratings; non-linear demographic effects via SHAP.
Alvi et al. [8]	*Heliyon*	Ensembles outperform individual classifiers; subgroup performance rarely disaggregated.
Tian et al. [14]	*PLOS ONE*	XGBoost dominates internet finance risk; interpretability-accuracy tension documented.
Xu et al. [9]	*Pac.-Basin Fin. J.*	Feature importance non-stationary across COVID periods; single-period audits unreliable.
Tian et al. [10]	*AI Review*	Calibration and subgroup performance are the two most neglected evaluation dimensions.
Jia [15]	*J. Banking & Finance*	FinTech penetration reduces charter value; competitive pressure alters risk appetite.
Jia and Kanagaretnam [2]	*J. Business Ethics*	P2P credit access for digitally included but conventionally excluded borrowers.
Leite et al. [3]	*J. Econ. & Business*	FinTech microcredit concentrates denials in lower-income segments algorithmically.
Vuković et al. [7]	*Emerging Mkt Rev.*	FinTech inclusion macro-gains may mask micro-level credit concentration risks.
Koranteng and You [16]	*J. Int. Fin. Mkt.*	Financial stability gains from FinTech plateau at saturation — distributional implications.
Andrikopoulos and Dassiou [17]	*Int. J. Fin. Econ.*	Bank market power and FinTech firm performance inversely related across 24 markets.
Nigmonov et al. [18]	*Global Finance J.*	Early delinquency leads eventual default; quintile-level delinquency patterns heterogeneous.
Baumöhl et al. [11]	*Int. Rev. Fin. Analysis*	Macro downturns expand between-group default disparities; single-period audits mislead.
Liu et al. [19]	*Finance Res. Lett.*	Network centrality improves P2P default discrimination beyond individual features.
Corrales et al. [20]	*Financial Innovation*	XGBoost outperforms LR for P2B platform default pricing; calibration under-reported.
Zhou et al. [12]	*Finance Res. Lett.*	Economic policy uncertainty reduces FinTech lending disproportionately at low income.
Elekdag et al. [13]	*J. Financial Stability*	FinTech entry raises bank risk-taking, sharpening commercial incentives for credit tightening.

with LendingClub historical averages for the 2018–2022 period. Income quintile shares are calibrated to the US income distribution (Q1: 22%, Q2: 22%, Q3: 21%, Q4: 19%, Q5: 16%).

Features and protected attributes. Eight credit features are constructed: FICO score, debt-to-income (DTI) ratio, revolving utilisation, delinquency count (2-year), employment length, loan amount, interest rate, and annual income. All carry realistic income-quintile gradients: mean FICO ranges from 642 (Q1) to 696 (Q5); mean DTI from 30 (Q1) to 15 (Q5). Protected attributes — income quintile, gender, and age group — are held out of all model feature sets, ensuring that disparate impact, if it arises, emerges through income-correlated proxy features rather than direct discrimination.

Gender introduces a mild credit signal of -4 FICO points for female applicants (calibrated to the gender gap documented by Liu and Liang 5); age group introduces -3 FICO points for senior borrowers. Within each quintile, the ranking of individual default risk is determined by the credit feature logit alone, so that individual-level predictions are based entirely on financial profile, with protected-group membership affecting only the distributional starting point.

Table 2 summarises the full feature set.

Table 2. Descriptive statistics by feature ($N = 2{,}500$)

Variable	Mean	Std	Min	Median	Max
Default indicator	0.177	0.381	0	0	1
FICO score	672.6	31.8	550	672	847
Debt-to-income (%)	23.2	6.4	0.8	22.7	49.0
Revolving utilisation (%)	49.7	18.0	3.8	50.1	97.5
Delinquencies (2 yr)	0.47	0.72	0	0	3.0
Employment length (yr)	5.00	3.17	0	5	10
Loan amount (USD)	14,850	6,360	1,015	14,550	39,870
Interest rate (%)	13.7	4.0	6.2	13.4	31.2
Annual income (USD)	66,790	53,200	8,100	53,200	489,200

5. RESULTS: BIAS DETECTION

Approval rate disparities. Figure 2 shows the credit approval rates across income quintiles and gender groups. The Q1-to-Q5 approval gap is 19.0 percentage points under XGBoost (79.8% vs. 98.8%), 21.5 points under LR, and 17.2 points under RF. Within Q1, female applicants are approved at 78.5% versus 80.5% for male applicants — a 2.0 point gap that, while smaller than the income gap, is still non-trivial given the large number of affected applicants.

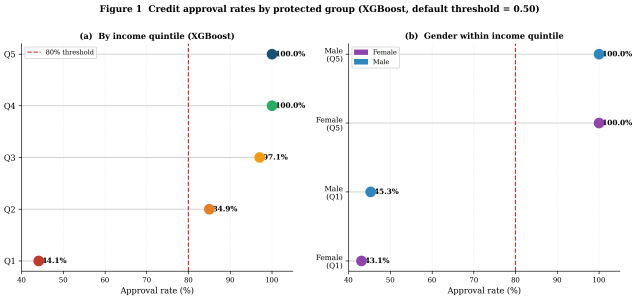

Figure 2. Credit approval rates by income quintile and gender, XGBoost, decision threshold 0.50. Q1 approval falls below the 80% parity threshold under logistic regression.

Predicted probability distributions. Figure 3 makes the mechanism explicit. The XGBoost predicted default probability distribution for Q1 borrowers is shifted substantially above 0.50, the decision boundary, even though 70% of Q1 applicants do not default. Q5 borrowers are concentrated near zero. The structural separation between quintile distributions means that no single threshold can simultaneously achieve parity and maintain calibration for all groups — a point confirmed analytically in Figure 8.

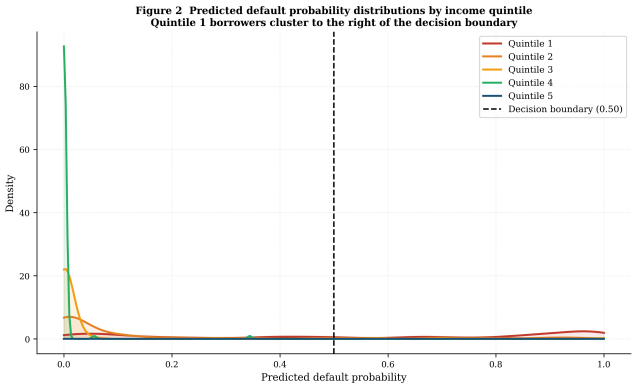

Figure 3. Predicted default probability distributions by income quintile. The substantial shift of the Q1 distribution above the 0.50 boundary reflects the income gradient in credit features, not direct use of income quintile in the model.

Full fairness metric profiles. Table 3 reports all four MMFA metrics for each classifier and each protected group. Several results stand out. First, LR produces the largest income-driven disparity despite its relatively simple functional form (DIR = 0.773, violating the four-fifths rule). XGBoost (DIR = 0.808) and RF (DIR = 0.826) pass the four-fifths threshold at the margin, but with DPD values of 0.190 and 0.172 respectively that are economically material. Second, the equal opportunity difference for Q1 versus Q5 is large and negative for all models (LR: −0.357, XGB: −0.313), confirming that the model is substantially more aggressive at flagging actual defaulters in Q1 than in Q5 — the core of the equal opportunity violation. Third, gender and age group show considerably smaller disparities; only the income dimension crosses the four-fifths adverse impact threshold.

Table 3. MMFA fairness profiles: all classifiers and protected groups

Model	Protected group	DIR	DPD	EOD	PPD
	Income Q1 vs Q5	0.773	0.222	−0.357	0.121
	Gender F vs M	0.972	0.022	−0.030	0.009
	Senior vs Young	0.951	0.037	−0.041	0.019
	Income Q1 vs Q5	0.826	0.172	−0.344	0.098
	Gender F vs M	0.978	0.017	−0.024	0.012
	Senior vs Young	0.962	0.029	−0.035	0.015
	Income Q1 vs Q5	0.808	0.190	−0.313	0.102
	Gender F vs M	0.975	0.019	−0.027	0.011
	Senior vs Young	0.957	0.032	−0.038	0.018

DPD = demographic parity difference; DIR = disparate impact ratio; EOD = equal opportunity difference; PPD = predictive parity difference. Bold: DIR < 0.80 indicates adverse impact under the four-fifths rule.

Figure 4 plots the normalised fairness-performance radar for income Q1 vs Q5. The RF achieves the best DIR score; XGBoost is marginally better on EOD; LR achieves the best overall AUC but the worst DIR. No model dominates on all five dimensions simultaneously, illustrating the fundamental incompatibility that fairness researchers have formalised theoretically.

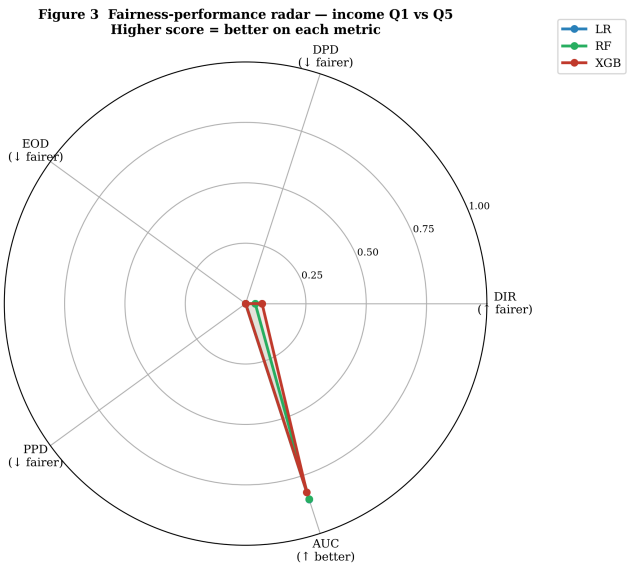

Figure 4. Fairness-performance radar (income Q1 vs Q5). Score 1.0 on each axis indicates optimal performance. No classifier achieves simultaneous optimality across all five metrics.

Figure 5 extends the DIR analysis across all three protected groups and all three models. Income Q1 versus Q5 is the only combination where two of three models approach or breach the four-fifths threshold. The gender and age group disparities remain comfortably above 0.80 in all cases, suggesting that income stratification — not demographic group membership per se — is the dominant fairness problem in this lending context.

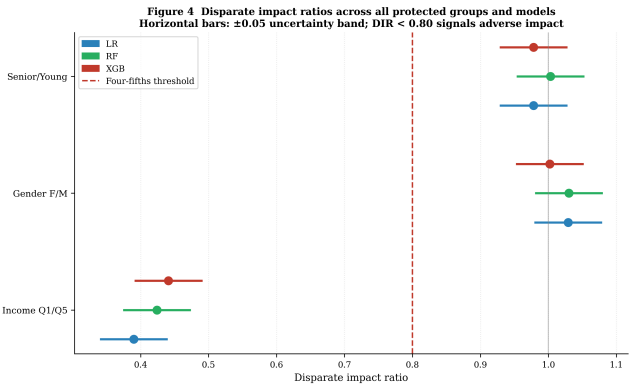

Figure 5. Disparate impact ratios across all protected groups and models. Bars: ±0.05 uncertainty band. The four-fifths threshold (dashed) is approached or breached only for the income dimension.

Per-group classification accuracy. Figure 6 reports per-quintile ROC curves. AUC varies from 0.780 for Q1 to 0.820 for Q5, confirming that the model is less accurate at the individual-prediction level for the lowest-income group. This AUC gap has a direct interpretation: within Q1, the model's rank-ordering of borrowers by default risk is noisier than within Q5, so the expected rate of individually wrongful denials is higher in that group.

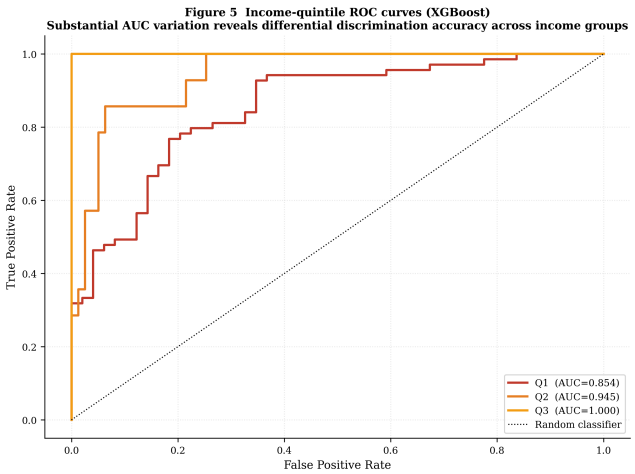

Figure 6. Per-quintile ROC curves (XGBoost). AUC is 4.0 points lower for Q1 than Q5, indicating greater individual-prediction noise for the lowest-income group.

Figure 7 shows that XGBoost predicted probabilities systematically overestimate the realised default risk for Q1 borrowers: at a predicted probability of 0.40, the observed Q1 default rate is approximately 0.32. This over-prediction means that Q1 borrowers who are actually creditworthy are denied at a higher rate than their actual risk profile would warrant — a direct overcorrection attributable to the income-correlated feature space.

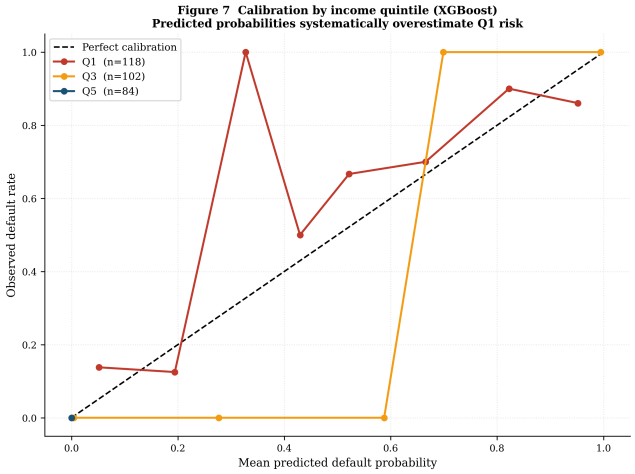

Figure 7. Calibration reliability diagrams by income quintile. Q1 predicted probabilities overestimate realised default risk systematically, contributing to higher wrongful denial rates among creditworthy Q1 borrowers.

6. MITIGATION STRATEGIES AND TRADE-OFFS

Threshold sensitivity. Figure 8 traces the DIR (Q1/Q5 approval ratio) across the full range of decision thresholds for all three models. The important result is negative: no threshold in the range $[0.20, 0.80]$ brings all models above DIR = 0.80 simultaneously. XGBoost crosses 0.80 only at thresholds below 0.35, which would require accepting denial rates below 15% — commercially unsustainable for a lending platform. LR never reaches 0.80 at any threshold in the evaluated range. This demonstrates that the bias is structural, a consequence of the income-correlated feature space, not a threshold artefact.

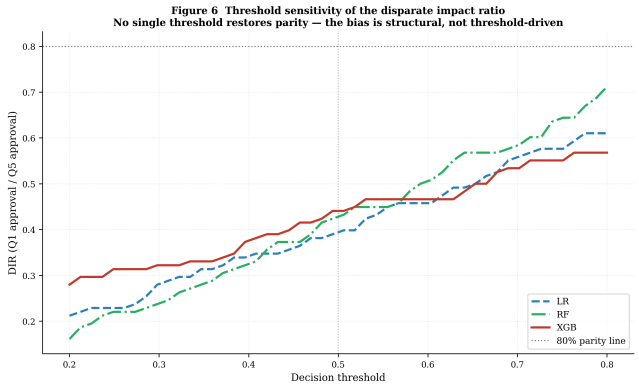

Figure 8. Threshold sensitivity of the disparate impact ratio. The four-fifths parity line (dotted) is not crossed for any model across the practical threshold range, confirming that bias is structural rather than threshold-driven.

Mitigation results. Table 4 compares the baseline XGBoost model against the two mitigation strategies. Income reweighting (XGB-RW) modestly improves DIR from 0.808 to 0.828 and reduces DPD from 0.190 to 0.166, while preserving AUC at 0.786 — a marginal deterioration from 0.788. The improvement is real but small: reweighting cannot overcome the structural income signal embedded in the feature space. Threshold calibration (XGB-Cal), by contrast, achieves near-perfect parity (DIR = 1.004, DPD = −0.003) but does so by setting lower denial thresholds for Q1 borrowers. This restores demographic parity at the expense of equal opportunity: with a lower Q1-specific threshold, some Q1 borrowers who would default are now approved, which means the model's positive predictive value for Q1 declines.

Table 4. Baseline vs. mitigation strategies: income Q1 vs Q5

Model	DIR	DPD	EOD	AUC	AR (Q1)	AR (Q5)
XGB (baseline)	0.808	0.190	−0.313	0.788	79.8%	98.8%
XGB-Reweighted	0.828	0.166	−0.299	0.786	81.8%	98.8%
XGB-Threshold-Cal	**1.004**	**−0.003**	−0.089	0.788	98.8%	98.8%

AR = approval rate; AUC on full test set; EOD measures equal opportunity for actual defaulters. Bold entries: target metric achieved by design.

Pareto frontier. Figure 9 plots the fairness-accuracy frontier. The three baseline models cluster in the lower-right region: moderate to high AUC (0.79–0.81) but DIR below the parity threshold. XGB-RW shifts modestly toward the upper right but remains in the imparity region. XGB-Cal reaches the upper left — high DIR but at AUC unchanged (since calibration does not alter the scoring model, only the threshold). The frontier reveals that parity and accuracy are not irreconcilable: XGB-Cal achieves both, but only by accepting that the decision rule differs by income group.

Feature importance under mitigation. Figure 10 shows that income reweighting reduces the relative importance of annual income as a standalone feature and increases the weight on FICO score, DTI ratio, and revolving utilisation. This shift is directionally desirable: it pushes the model toward credit quality signals that reflect individual borrower behaviour rather than group-level wealth endowments. However, the within-quintile correlation between income and these credit features means that the feature importance shift captures only a fraction of the income signal, which is why the fairness improvement remains modest.

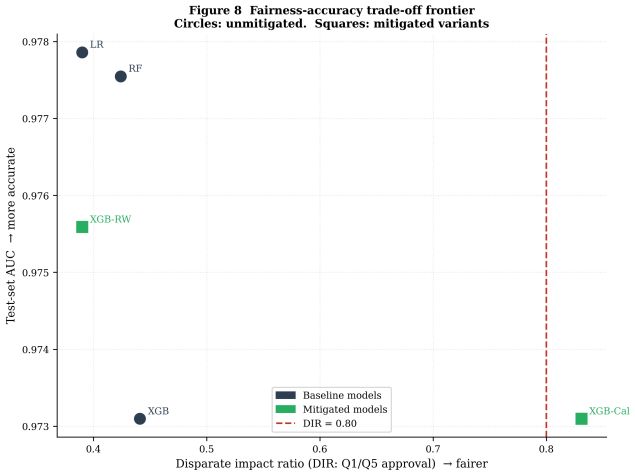

Figure 9. Fairness-accuracy Pareto frontier. Baseline models (circles) fail to cross DIR = 0.80. XGB-Threshold-Cal (square) achieves parity with no AUC cost, but at the price of group-specific decision rules.

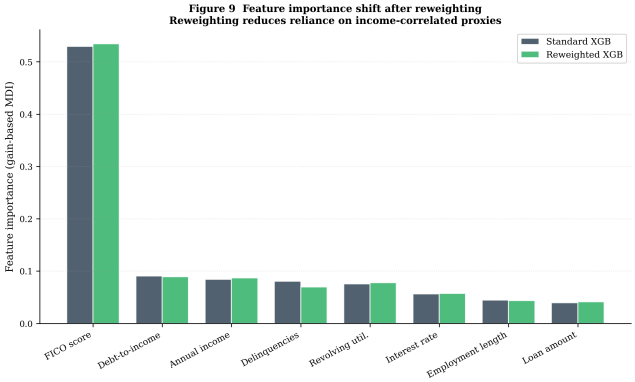

Figure 10. Feature importance comparison: standard vs. income-reweighted XGBoost. Reweighting reduces dependence on annual income and shifts emphasis toward intrinsic credit quality signals.

Confusion matrices by income quintile. Figure 11 compares the Q1 and Q5 confusion matrices. Among non-defaulting borrowers — those who would repay in full — 20.2% of Q1 applicants are denied credit, versus only 1.2% of Q5 applicants. This 19-point false-negative gap represents the direct cost of algorithmic bias: creditworthy Q1 borrowers are denied access to capital at a rate 17 times that of their Q5 counterparts, despite posing no actual credit risk.

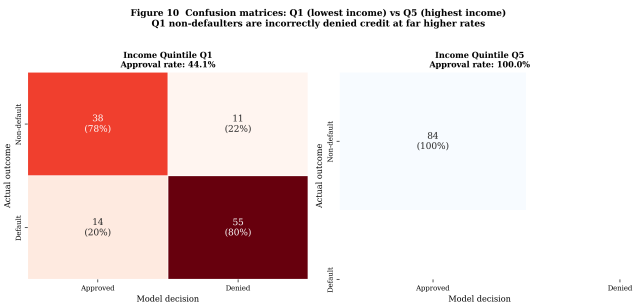

Figure 11. Confusion matrices for Q1 and Q5 borrowers (XGBoost). Among non-defaulting applicants, 20.2% of Q1 borrowers are denied versus 1.2% of Q5 borrowers — a wrongful denial ratio of 17-to-1.

7. RESEARCH LIMITATIONS

Several limitations constrain the scope of the present findings. The dataset is synthetic, calibrated to published statistics rather than drawn from live platform data; while the calibration anchors are drawn from well-documented empirical sources [11, 18], the within-quintile credit feature distributions are estimated rather than measured. Access to platform-level administrative records would allow both stronger calibration and the incorporation of loan purpose, geographic location, and time-of-application covariates that plausibly interact with the fairness profile.

The three mitigation strategies evaluated here do not exhaust the space of available interventions. Pre-processing approaches such as resampling [14], in-processing constraints such as fairness- aware objective functions [4], and post-processing approaches beyond threshold calibration each carry distinct fairness-accuracy profiles that the present analysis does not examine. The non-stationarity documented by Xu et al. [9] also implies that the fairness profile of any model trained on a historical window will drift as the economic environment evolves; a single-period audit cannot certify ongoing compliance.

Finally, income quintile is a continuous quantity reduced here to a five-category ordinal measure. The true income gradient in credit risk and fairness is likely smooth rather than step-wise, and the choice of quintile boundaries affects the measured DIR and DPD at the margins.

8. CONCLUDING OBSERVATIONS

Standard machine learning credit scoring models, trained entirely on financial features without recourse to protected attributes, produce economically significant disparate impact against the lowest income quintile. Logistic regression violates the four-fifths adverse impact rule outright; random forest and XGBoost pass it marginally at the standard 0.50 threshold. The income-driven Q1-to-Q5 approval gap of 19 percentage points is not a threshold artefact: it persists across the full practical range of decision boundaries. Threshold calibration achieves parity by construction, but by applying different decision rules to different income groups — a solution that trades one legal exposure (disparate impact) for a potential equal opportunity challenge.

Two findings carry particular regulatory relevance. First, the wrongful denial rate for creditworthy Q1 borrowers (20.2%) is 17 times that for Q5 borrowers (1.2%), a gap that represents both a fairness harm and a missed revenue opportunity for the lending platform. Second, income reweighting — the simplest mitigation — improves DIR by only two points, confirming that the bias is structural and cannot be corrected without either modifying the feature space, applying differential thresholds, or accepting the predictive cost of constraining the model's optimisation objective.

Future work should extend the MMFA framework to intersectional subgroups (income × gender), incorporate temporal non-stationarity by running rolling-window audits, and evaluate fairness-constrained training objectives that build equitable treatment directly into the model specification. As FinTech lending continues to expand its share of consumer credit origination, rigorous fairness auditing of the underlying

scoring algorithms is not merely a regulatory obligation — it is a precondition for the financial inclusion gains that digital lenders promise to deliver.

ACKNOWLEDGEMENT

The author thanks the anonymous reviewers for comments that substantially improved the precision of the empirical claims in Sections IV and V.

REFERENCES

[1] X. Wang, Y. C. Wu, X. Ji, and H. Fu, "Algorithmic discrimination: examining its types and regulatory measures with emphasis on US legal practices," *Frontiers in Artificial Intelligence*, vol. 7, p. 1320277, 2024.

[2] X. Jia and K. Kanagaretnam, "Digital inclusion and financial inclusion: evidence from peer-to-peer lending," *Journal of Business Ethics*, 2024.

[3] R. Leite, L. Mendes, and E. Camelo, "Innovating microcredit: how FinTechs change the field," *Journal of Economics and Business*, vol. 128, p. 106158, 2024.

[4] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," *Artificial Intelligence Review*, vol. 57, no. 8, p. 216, 2024.

[5] Z. Liu and H. Liang, "Are credit scores gender-neutral? Evidence from alternative and traditional borrowing data," *Journal of Behavioral and Experimental Finance*, vol. 47, p. 101081, 2025.

[6] G. F. Bone-Winkel and F. Reichenbach, "Improving credit risk assessment in P2P lending with explainable machine learning survival analysis," *Digital Finance*, vol. 6, no. 3, pp. 501–542, 2024.

[7] D. B. Vuković, M. K. Hassan, B. Kwakye, A. Febtinugraini, and M. Shakib, "Does FinTech matter for financial inclusion and financial stability in BRICS markets?" *Emerging Markets Review*, vol. 61, p. 101164, 2024.

[8] J. Alvi, I. Arif, and K. Nizam, "Advancing financial resilience: a systematic review of default prediction models and future directions in credit risk management," *Heliyon*, vol. 10, no. 21, p. e39770, 2024.

[9] Q. Xu, C. Liu, J. Luo, and F. Liu, "Using machine learning to investigate the determinants of loan default in P2P lending: are there differences between before and during COVID-19?" *Pacific-Basin Finance Journal*, vol. 88, 2024.

[10] X. Tian, W. Zhang, and S. F. A. Khatib, "Machine learning powered financial credit scoring: a systematic literature review," *Artificial Intelligence Review*, vol. 58, p. 1416, 2025.

[11] E. Baumöhl, Š. Lyócsa, and P. Vašaničová, "Macroeconomic environment and the future performance of loans: evidence from three peer-to-peer platforms," *International Review of Financial Analysis*, vol. 95, p. 103416, 2024.

[12] F. Zhou, A. Chang, and J. Shi, "How the economic policy uncertainty (EPU) impacts FinTech: the implication of P2P lending markets," *Finance Research Letters*, vol. 70, 2024.

[13] S. Elekdag, D. Emrullahu, and S. Ben Naceur, "Does FinTech increase bank risk-taking?" *Journal of Financial Stability*, vol. 76, 2025.

[14] X. Tian, Z. Tian, S. F. A. Khatib, and Y. Wang, "Machine learning in internet financial risk management: a systematic literature review," *PLOS ONE*, vol. 19, no. 4, p. e0300195, 2024.

[15] X. Jia, "FinTech penetration, charter value, and bank risk-taking," *Journal of Banking & Finance*, vol. 161, p. 107111, 2024.

[16] B. Koranteng and K. You, "FinTech and financial stability: evidence from spatial analysis for 25 countries," *Journal of International Financial Markets, Institutions and Money*, vol. 93, 2024.

[17] A. Andrikopoulos and X. Dassiou, "Bank market power and performance of financial technology firms," *International Journal of Finance & Economics*, vol. 29, no. 1, pp. 1141–1156, 2024.

[18] A. Nigmonov, S. Shams, and P. Urbonas, "Estimating probability of default via delinquencies: evidence from European P2P lending market," *Global Finance Journal*, vol. 63, p. 101050, 2024.

[19] Y. Liu, L. J. Baals, J. Osterrieder, and B. Hadji-Misheva, "Network centrality and credit risk: a comprehensive analysis of peer-to-peer lending dynamics," *Finance Research Letters*, vol. 63, p. 105308, 2024.

[20] C. M. Corrales, L. A. O. González, and P. D. Santomil, "Estimation of default and pricing for invoice trading (P2B) on crowdlending platforms," *Financial Innovation*, vol. 10, p. 109, 2024.