World Journal of Sociology and Law
Print ISSN: 2960-0294
Online ISSN: 2960-0308
DOI: https://doi.org/10.61784/wjsl4034

ALGORITHMIC BIAS AND DISCRIMINATION IN AI-DRIVEN LENDING: A RASHOMON-BASED AUDIT OF LESS DISCRIMINATORY ALTERNATIVES AND ADVERSE-ACTION ACCOUNTABILITY

XiaoGuo Chen
Georgetown University Law Center, Washington, DC 20001, USA.

Abstract: AI-driven credit scoring is governed by rules that ask not only whether outcomes differ across protected groups, but also whether an equally effective, less discriminatory alternative (LDA) exists and whether the reasons given for adverse decisions are specific and accurate. This paper proposes DIRA, a Disparate-Impact Rashomon Audit that maps U.S., EU and Chinese requirements onto four measurable stages: prima facie disparity, proxy leakage, Rashomon-based LDA search and adverse-action multiplicity. Its R-LDA and R-LDA-E procedures search near-optimal gradient-boosting models enriched with proxy-attenuated variants, use protected attributes only for auditing, and confirm every alternative on held-out data. Experiments on German Credit, Taiwan Default and Home Credit (30 random splits, 3,054 credit models) show that excluding sex and age does not hide them: age is recoverable from the remaining inputs with an AUC of up to 0.988. Within an AUC tolerance of 0.01, the test worst-case adverse impact ratio (AIR) of equally accurate Home Credit models ranges from 0.540 to 0.680. R-LDA-E raises the worst-case AIR from 0.606 to 0.647 on Home Credit and by 0.026 on German Credit with AUC changes of at most 0.0011, whereas naive selection overstates the attainable gain by a factor of up to 4.1. Equally accurate models disagree on up to 22.3% of decisions, more often for applicants under 25, and change the principal denial reason for 24–57% of denied applicants. Business-necessity defences, LDA claims and adverse-action explanations should therefore be assessed against the full set of equally effective models rather than a single scorecard.
Keywords: Algorithmic fairness; Credit scoring; Disparate impact; Less discriminatory alternative; Model multiplicity; Explainability; Fair lending law; AI regulation

1 INTRODUCTION

Machine learning (ML) now sits at the core of consumer credit underwriting. Gradient-boosted trees trained on hundreds of bureau, application and alternative-data variables promise sharper risk ranking and wider access for thin-file borrowers. The same flexibility, however, redistributes credit in ways that are hard to see. Fuster *et al.* show on U.S. mortgage data that the gains from ML default models accrue unevenly across racial groups and that rate disparities widen [1], chiefly because flexible models exploit structure that simpler scorecards ignore. Bartlett *et al.* find that algorithmic lenders still charge Latinx and African-American borrowers higher rates [2], although less than face-to-face lenders. Algorithmic bias in lending is therefore not hypothetical; it is a measurable property of deployed scoring pipelines.

The legal system addresses this risk through several overlapping instruments. In the United States, the Equal Credit Opportunity Act (ECOA) and Regulation B prohibit discrimination on the basis of sex, marital status, age and other characteristics, and the regulation, as in force in 2023, recognises an effects test for practices with a disproportionate adverse impact [3]. The Supreme Court confirmed disparate-impact liability under the Fair Housing Act in *Inclusive Communities* [4], whose burden-shifting structure ends with the question of whether a less discriminatory alternative (LDA) could serve the lender's legitimate interest. Regulation B further requires specific principal reasons for adverse action, and the Consumer Financial Protection Bureau (CFPB) has stated that model complexity does not excuse inaccurate or generic reasons [5-6]. In the European Union, the proposed AI Act classifies creditworthiness evaluation as high-risk [7]; the Court of Justice held in *SCHUFA* that a credit score playing a determining role is an automated decision under Article 22 GDPR [8-9]; and the new Consumer Credit Directive grants consumers explanation and human-intervention rights for automated creditworthiness assessments [10]. China's Personal Information Protection Law (PIPL) requires automated decisions to be transparent and fair and prohibits unreasonable differential treatment in transaction terms [11].

Fair-ML practice and these doctrines are only partially aligned [12-13]. Three gaps stand out. First, many mitigation methods consume the protected attribute at decision time, which collides with disparate-treatment prohibitions. Second, audits typically evaluate a single deployed model, whereas the doctrine asks whether an *equally effective* alternative with smaller disparity exists. Model multiplicity implies that such alternatives are the rule rather than the exception [14-16], so the choice among equally accurate models is itself a legally consequential act. Third, explanation duties presume that a model's principal reasons are well defined, yet under multiplicity the reasons given to an applicant may depend on arbitrary development choices. Removing protected attributes does not close these gaps, because flexible models reconstruct protected status from proxies [17-18].

© By the Author(s) 2023, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

This paper proposes DIRA, a Disparate-Impact Rashomon Audit that turns the burden-shifting structure and the adverse-action duty into measurable, reproducible tests. Its contributions are as follows.

• A doctrine-to-metric mapping that connects U.S., EU and Chinese rules on credit discrimination and automated decisions to concrete audit quantities (Section 3).

• R-LDA and its ensemble variant R-LDA-E, which search a Rashomon set enriched with *proxy-attenuated* models, use protected attributes only at the audit stage, and confirm every candidate alternative on held-out data (Section 4).

• Adverse-action multiplicity measures showing that both credit decisions and principal reasons change across equally accurate models, with a disproportionate burden on young applicants (Section 6).

• An empirical study on three public credit datasets with 30 random splits and 3,054 trained models, comparing DIRA with five alternatives, including FairGBM [19], under a legal-admissibility taxonomy.

2 RELATED WORK

2.1 Fairness in Credit Scoring

Hardt *et al.* formalised equal opportunity, illustrated it on FICO scores and showed that group-specific thresholds can post-process any score into compliance with a chosen criterion [20]. Kamiran and Calders introduced reweighing as a pre-processing remedy [21]. Kozodoi *et al.* benchmarked fairness processors on seven credit datasets and analysed their profit implications [22], recommending separation-based criteria for scorecards. FairGBM embeds smooth error-rate constraints in gradient boosting through a proxy-Lagrangian formulation and is a strong in-processing method for tabular financial data [19]. These methods differ sharply in *where* the protected attribute enters the pipeline, a distinction with direct legal consequences that is rarely evaluated together with accuracy and fairness.

2.2 Law and Algorithmic Discrimination

Barocas and Selbst showed that data mining can generate disparate impact that existing antidiscrimination law struggles to reach [23]. Prince and Schwarcz argued that AI systems deprived of protected attributes rationally locate proxies [18], and Gillis demonstrated in credit pricing that excluding inputs neither prevents disparities nor guarantees fairness [17], calling for outcome-focused analysis. Kumar *et al.* aligned fair-ML research with U.S. fair-lending regulation [12], Hall *et al.* described compliance practice built on the adverse impact ratio (AIR) and the standardised mean difference (SMD) [24], and Wachter *et al.* analysed why statistical fairness cannot be mechanically transposed into EU non-discrimination law [13].

2.3 Model Multiplicity and Explanation

Marx *et al.* defined ambiguity and discrepancy to measure predictive multiplicity [14]; Watson-Daniels *et al.* extended the analysis to probabilistic classifiers and observed that multiplicity can weigh more heavily on minority groups [16]; Black *et al.* discussed how multiplicity both enables LDA search and raises concerns about arbitrariness [15]. TreeSHAP is the standard route from tree ensembles to per-applicant reason codes [25]. DIRA joins these lines: it operationalises LDA search with protected-attribute-blind candidate generation and out-of-sample confirmation, quantifies the optimism created by selecting among many candidates, and links multiplicity to adverse-action duties.

3 LEGAL AND ETHICAL FRAMEWORK

3.1 From Doctrine to Measurement

Table 1 summarises how DIRA translates legal requirements into audit quantities. Under U.S. law, using a prohibited basis in a credit decision is disparate treatment and is unlawful irrespective of motive; the main exception permits age in an empirically derived, demonstrably and statistically sound scoring system, provided that applicants aged 62 or older are not assigned a negative value [3]. Disparate-impact analysis proceeds in three steps: a prima facie showing that a facially neutral practice produces a disproportionate adverse effect; a justification showing that the practice serves a legitimate business need; and a rebuttal showing that a less discriminatory alternative would serve that need equally well [4]. EU non-discrimination law reaches the same structure through objective justification and proportionality, whose necessity limb asks whether less restrictive means are available [13]. The four-fifths rule provides a widely used screening threshold for the first step [26], which we complement with significance tests because the ratio alone ignores sampling error.

Table 1 Mapping of Legal Requirements to DIRA Audit Quantities

Legal source	Requirement	DIRA stage and metric
ECOA/Reg. B §1002.4(a), §1002.6(b) [3]; EU direct discrimination [13]; PIPL Art. 24 [11]	No prohibited basis in decisions (age in scoring only with elderly safeguard)	Inputs exclude sex, age, marital status; PA-use tier (Table 5)
Reg. B §1002.6(a) effects test [3]; FHA [4]; four-fifths rule [26]	Prima facie disproportionate adverse effect	S1: AIR_a, AIR_{min}, z, SMD, AIR^q
Business necessity [4]; objective justification [13]	Legitimate, substantial interest	AUC and PDR of m^* ; ε-band
Less discriminatory alternative [4]; proportionality [13]	No equally effective, less discriminatory practice	S3: test ΔAIR_{min}, ΔAUC, ΔPDR, bootstrap CI

Legal source	Requirement	DIRA stage and metric
Proxy discrimination [17-18]; CCD Art. 18 [10]	Neutral inputs must not reconstruct protected status	S2: proxy AUC, PR_j, attenuation
Reg. B §1002.9(b)(2), [5-6]; GDPR Art. 22, SCHUFA [8-9]; CCD Art. 18 [10]; PIPL Art. 24 [11]	Specific, accurate, contestable reasons	S4: ambiguity, Jaccard J, top-1 flips
AI Act proposal, Annex III 5(b), Art. 10 [7]	Bias examination for high-risk scoring	Complete DIRA report (S1–S4)

3.2 Ethical Dimensions

Four ethical tensions shape the audit. *(i) Competing fairness criteria.* Selection-rate parity and equal opportunity among creditworthy applicants cannot generally hold together when default rates differ across groups [20]; DIRA therefore reports both AIR and AIR^q. *(ii) Arbitrariness.* When equally accurate models disagree, an applicant's outcome depends on unprincipled development choices such as a random seed, a procedural harm distinct from group disparity [15-16]. *(iii) Contestability.* Explanation rights under Regulation B, GDPR, the Consumer Credit Directive and PIPL presuppose reasons that reflect the applicant's circumstances rather than model idiosyncrasies. *(iv) Inclusion versus prudence.* Fairness interventions that raise the default rate of the approved portfolio can harm borrowers through over-indebtedness, a concern central to responsible-lending rules [10]. DIRA makes each tension explicit rather than resolving it implicitly through a single fairness metric.

4 METHODOLOGY: THE DIRA FRAMEWORK

Figure 1 outlines DIRA. Its four stages mirror the legal sequence of Table 1: S1 establishes whether a prima facie disparity exists, S2 explains *why* it survives the exclusion of protected attributes, S3 tests whether an equally effective but less discriminatory model exists, and S4 examines whether adverse-action reasons are stable enough to be specific and accurate.

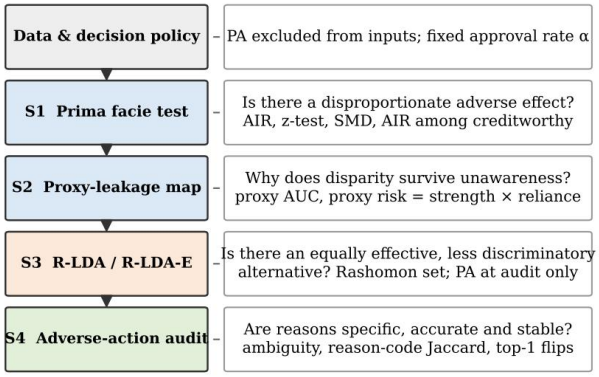

Figure 1 Overview of DIRA

Note: Each audit stage answers one step of the legal analysis; protected attributes (PA) are used only for auditing and model selection.

4.1 Setting and Decision Policy

Each applicant has features x (protected attributes excluded), a default label $y \in \{0, 1\}$, and protected attributes $a \in \mathcal{A} = \{\text{sex, age}\}$ that are visible only to the auditor. A model m outputs a risk score $s_m(x)$. The lender approves the fraction α of applicants with the lowest predicted risk, with the cut-off fixed on validation data:

$$\tau_\alpha = \text{Quantile}_\alpha\{s(x) : x \in \mathcal{D}_{val}\}, \quad \hat{y}(x) = \mathbb{1}[s(x) \leq \tau_\alpha] \tag{1}$$

Fixing α holds lending volume constant, so every alternative is compared at the same business scale and differences in disparity cannot be bought by simply approving more applicants.

4.2 S1: Prima Facie Disparity

For each protected attribute the adverse impact ratio compares approval rates across groups, and the audit focuses on the worst case:

$$\text{AIR}_a = \frac{\min_g \Pr[\hat{y} = 1 \mid a = g]}{\max_g \Pr[\hat{y} = 1 \mid a = g]}, \quad \text{AIR}_{\min} = \min_{a \in \mathcal{A}} \text{AIR}_a \tag{2}$$

Values below 0.8 fail the four-fifths screen [26]. DIRA also reports AIR^q, the same ratio restricted to creditworthy applicants ($y = 0$), which is an equal-opportunity analogue [20]; the two-proportion z statistic of the approval-rate gap; the SMD of risk scores [24]; and the portfolio default rate (PDR), the realised default rate among approved applicants.

4.3 S2: Proxy-Leakage Map

The joint proxy AUC measures how well a gradient-boosted classifier recovers a from x. At feature level, $AUC_j^{(a)}$ is the AUC of a single-feature model predicting a, and the decision reliance ρ_j is the normalised mean absolute TreeSHAP contribution φ_j of feature j in the incumbent model [25]. Their product defines the proxy risk

$$\mathrm{PR}_j = \left(2\,\mathrm{AUC}_j^{(a)} - 1\right) \cdot \frac{\rho_j}{\max_k \rho_k}, \quad \rho_j = \frac{\mathbb{E}\,|\varphi_j(x)|}{\sum_k \mathbb{E}\,|\varphi_k(x)|} \tag{3}$$

A feature with high PR_j is simultaneously informative about protected status and decisive for credit outcomes, which is precisely the configuration that produces proxy discrimination [18].

4.4 S3: Rashomon Search for a Less Discriminatory Alternative

The candidate pool $\mathcal{C} = \mathscr{H} \cup \mathcal{P}$ has two parts. $\mathscr{H}$ contains $M = 60$ LightGBM models with randomly sampled hyper-parameters (number of leaves, depth, learning rate, number of trees, minimum child size, row and column subsampling, L2 penalty, seed) [27]. The incumbent m^* is the most accurate member of $\mathscr{H}$ on validation data, i.e., the model a lender would deploy under business-as-usual development. $\mathcal{P}$ contains *proxy-attenuated* models: for nested sets S of the top-ranked proxy-risk features of each protected attribute (top-1, top-2, top-3 and the union of the top-2 sets) and attenuation levels $\gamma \in \{0, 0.25, 0.5\}$, the split gain of every feature in S is multiplied by γ during tree growth. Setting $\gamma = 0$ removes the feature; intermediate values discourage reliance on it while retaining residual predictive signal. Each (S, γ) pair is trained with the incumbent configuration and with one random configuration. No candidate uses a in training or prediction; a enters only the selection step, i.e., the audit. The Rashomon set collects all candidates whose validation AUC is within ε of the incumbent:

$$\mathcal{R}_\varepsilon = \{m \in \mathcal{C} : \mathrm{AUC}_{val}(m) \geq \mathrm{AUC}_{val}(m^*) - \varepsilon\} \tag{4}$$

R-LDA selects the member with the highest validation AIR_{min}; R-LDA-E averages the scores of the $K = 5$ members with the highest validation AIR_{min}, denoted $\mathcal{T}_K$:

$$m_{\mathrm{LDA}} = \arg\max_{m \in \mathcal{R}_\varepsilon} \mathrm{AIR}_{\min}^{val}(m), \quad s_{\mathrm{E}}(x) = \frac{1}{K}\sum_{m \in \mathcal{T}_K} s_m(x) \tag{5}$$

Selecting the fairest of many near-equivalent models rewards sampling noise as well as genuine fairness, so the apparent validation gain overstates the true gain. DIRA therefore reports only test-set improvements, computed on data never used for selection, and measures *selection-induced optimism* as the difference between the apparent and realised gains. Averaging the top-K members shrinks this optimism and stabilises decisions. Algorithm 1 summarises the procedure.

Algorithm 1. R-LDA and R-LDA-E

Input: D_{tr}, D_{val} (audit attributes a), approval rate α, tolerance ε, ensemble size K
1: Train M hyper-parameter variants $\mathscr{H}$ on D_{tr}; $m^* \leftarrow \mathrm{argmax}_{m \in \mathscr{H}} AUC_{val}(m)$
2: Compute PR_j on (D_{tr}, D_{val}) with m^* ; form nested proxy sets $\mathcal{S}$
3: for $S \in \mathcal{S}$, $\gamma \in \{0, 0.25, 0.5\}$: train attenuated models $\rightarrow \mathcal{P}$
4: $\mathscr{R}_\varepsilon \leftarrow \{m \in \mathscr{H} \cup \mathcal{P} : AUC_{val}(m) \geq AUC_{val}(m^*) - \varepsilon\}$
5: $m_{LDA} \leftarrow \mathrm{argmax}_{\mathscr{R}_\varepsilon} AIR^{val}_{min}$; $\mathcal{T}_K \leftarrow$ top-K of $\mathscr{R}_\varepsilon$ by AIR^{val}_{min}
6: $s_E \leftarrow$ mean of s_m, $m \in \mathcal{T}_K$; fix τ_α on D_{val} for every policy
7: return test-set ΔAIR_{min}, ΔAUC, ΔPDR versus m^* with bootstrap CIs

4.5 S4: Adverse-Action Multiplicity

Using the 20 most accurate members of $\mathscr{R}_\varepsilon$, DIRA computes *decision ambiguity*, the share of test applicants who receive conflicting decisions from equally accurate models [14], overall and per group. For each applicant i denied by m^*, the principal reasons $\mathcal{Q}_m(x_i)$ of model m are the (up to) four features with the largest positive TreeSHAP contributions toward default, matching the number of reasons customarily disclosed in adverse-action notices. Reason stability averages the Jaccard agreement over the set $\mathscr{M}_i$ of members that also deny i:

$$J_i = \frac{1}{|\mathcal{M}_i|}\sum_{m \in \mathcal{M}_i} \frac{|\mathcal{Q}_{m^*}(x_i) \cap \mathcal{Q}_m(x_i)|}{|\mathcal{Q}_{m^*}(x_i) \cup \mathcal{Q}_m(x_i)|} \tag{6}$$

A *top-1 flip* occurs when at least one member of $\mathscr{M}_i$ reports a different principal reason than m^*. Low J_i or a top-1 flip means that the reason an applicant receives is partly an artefact of model choice, which undermines the specificity and accuracy that Regulation B, GDPR Article 22 and PIPL Article 24 demand.

4.6 Baselines and Admissibility Tiers

DIRA is compared with five alternatives built on the incumbent's hyper-parameters: *Aware*, which adds the protected attributes as inputs; *Reweighing* [21]; *FairGBM* [19], constraining the false-positive rate of default predictions, i.e., false denials of creditworthy applicants, to be equal across the four sex–age groups; *Proxy removal*, which deletes the union of the top-3 proxy features of both attributes, the input-scrutiny remedy criticised in [17]; and *Group thresholds*, which sets group-specific cut-offs that equalise approval rates [20]. We classify each method by where the protected attribute enters: tier T0, audit only (Proxy removal, R-LDA, R-LDA-E); tier T1, training only (Reweighing, FairGBM); tier T2, decision time (Aware, Group thresholds). Legal risk under disparate-treatment rules rises from T0 to T2.

5 EXPERIMENTAL SETUP

Table 2 lists the datasets, all publicly released before 2023. German Credit contains 1,000 consumer loans [28]; Taiwan Default contains 30,000 credit-card accounts [29]; Home Credit contains 307,511 instalment-loan applications [30], from which we draw a random sample of 50,000. The audited attributes are sex (female versus male) and age (younger than 25 versus 25 or older). Model inputs exclude sex, age and marital status in all datasets and foreign-worker status in German Credit; Taiwan Default adds a utilisation and a repayment ratio, and Home Credit adds four affordability ratios. Splits are stratified by label, sex and age into 60/20/20 train, validation and test sets and repeated 20 times for German Credit and five times for the larger datasets. We use $\alpha = 0.8$, $\varepsilon = 0.01$ AUC, $K = 5$ and 300 bootstrap resamples of the test set; reason codes are computed for a random subset of at most 1,500 denied applicants per split. In total, 3,054 credit-scoring models and 1,400 auxiliary proxy models were trained.

Table 2 Datasets and Audited Groups

Dataset	n	Inputs	Default	Female	Age < 25	Splits
German	1,000	17	0.300	0.310	0.149	20
Taiwan	30,000	22	0.221	0.604	0.089	5
Home Credit†	50,000	44	0.082	0.664	0.038	5

Note: †Random sample of 50,000 of 307,511 applications. Shares are computed over the full dataset.

6 RESULTS

6.1 Prima Facie Disparity and Proxy Leakage

Table 3 reports the S1 audit of the incumbent models. They rank risk well (AUC 0.788, 0.784 and 0.750) but approve young applicants markedly less often. On Home Credit the worst-case AIR is 0.606 and fails the four-fifths screen in 100% of splits; on German Credit the age AIR is 0.870 and falls below 0.8 in 30% of splits. Taiwan Default never fails the screen (age AIR 0.909), yet its approval gap is statistically significant ($|z| > 1.96$) in 100% of splits, which shows that the ratio screen and significance testing answer different legal questions [24, 26]. Sex disparities are smaller (AIR between 0.890 and 0.958); the disadvantaged group is female applicants in German Credit and male applicants in the two larger datasets. On Home Credit, creditworthy young applicants are approved at only 0.631 times the rate of creditworthy older applicants (AIR^q), so the gap is not explained by default risk alone.

Table 3 Prima Facie Audit of the Incumbent Model (Test Set, Mean ± SD)

Metric	German	Taiwan	Home Credit
AUC	0.788 ± 0.033	0.784 ± 0.004	0.750 ± 0.008
PDR	0.205 ± 0.020	0.134 ± 0.002	0.050 ± 0.001
AIR_{sex}	0.958 ± 0.037	0.951 ± 0.008	0.890 ± 0.012
AIR_{age}	0.870 ± 0.095	0.909 ± 0.009	0.606 ± 0.015
AIR^q_{age}	0.904 ± 0.072	0.955 ± 0.014	0.631 ± 0.017
z_{age}	−1.23 ± 1.01	−4.02 ± 0.39	−15.20 ± 0.62
SMD_{age}	0.33 ± 0.19	0.29 ± 0.03	0.92 ± 0.05
4/5 violations (age)	30%	0%	100%
Proxy AUC (sex)	0.672	0.605	0.885
Proxy AUC (age)	0.742	0.810	0.988
Aware AIR_{age}	0.838	0.913	0.610

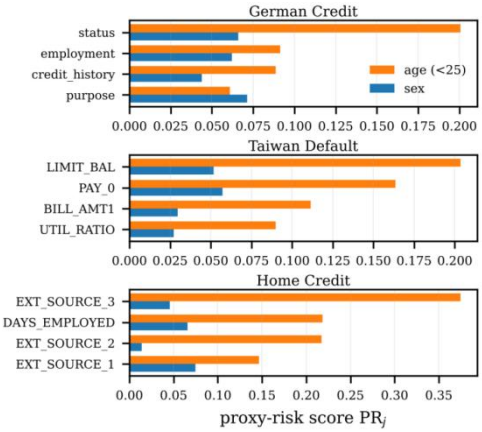

Figure 2 Proxy-Risk Map (S2): Top Features by Proxy Risk PRj, Averaged over Splits
Note: Long bars mark inputs that both reveal the protected attribute and drive credit decisions.

Excluding sex and age does not remove them from the data. A gradient-boosted classifier recovers age from the remaining inputs with AUC 0.742, 0.810 and 0.988, and sex with 0.672, 0.605 and 0.885. Figure 2 locates the leakage that matters for decisions: the credit limit in Taiwan Default (single-feature age AUC 0.801), external bureau scores and employment

duration in Home Credit (DAYS_EMPLOYED: 0.805), and checking-account status and employment length in German Credit. Adding the protected attributes as inputs does not help either: the age AIR of the Aware model is 0.838 on German Credit and 0.610 on Home Credit, against 0.870 and 0.606 for the incumbent, while the scorecard becomes explicit disparate treatment.

6.2 Disparity Multiplicity and the Search for an LDA

Equally accurate models are not equally fair. The ε = 0.01 band contains on average 22.1, 67.6 and 77.4 candidates on German Credit, Taiwan Default and Home Credit. Their test AUCs lie within 0.038, 0.014 and 0.015 of one another, yet their test worst-case AIR spans 0.805–0.948, 0.890–0.926 and 0.540–0.680 (Figure 3). The disparity of the incumbent is therefore partly an artefact of development choices such as the hyper-parameter search, rather than a necessary consequence of predicting default.

Table 4 summarises the held-out effect of adopting an alternative. On Home Credit, R-LDA raises the worst-case AIR by +0.062 and R-LDA-E by +0.041, in 100% and 100% of splits. The ensemble achieves this with an AUC change of −0.0011, a bootstrap AUC interval that contains zero in 100% of splits and an essentially unchanged portfolio default rate (+0.0001). On German Credit the gains are +0.028 and +0.026, the latter at no AUC cost (+0.0006). On Taiwan Default no material alternative exists at α = 0.8 (+0.004 and −0.005).

Selecting among many near-equivalent candidates produces substantial optimism. The apparent validation gain of R-LDA is +0.067 on German Credit, 2.4 times its realised test gain (+0.028), and on Taiwan Default an apparent gain of +0.016 shrinks to +0.004. Averaging curbs this effect: R-LDA-E has apparent and realised gains of +0.038 and +0.026 on German Credit, and on Home Credit its realised gain (+0.041) exceeds the apparent one (+0.030). Proxy-attenuated candidates supply 70%, 40% and 20% of the R-LDA selections; on Home Credit, hyper-parameter variation alone already spans a wide range of disparity.

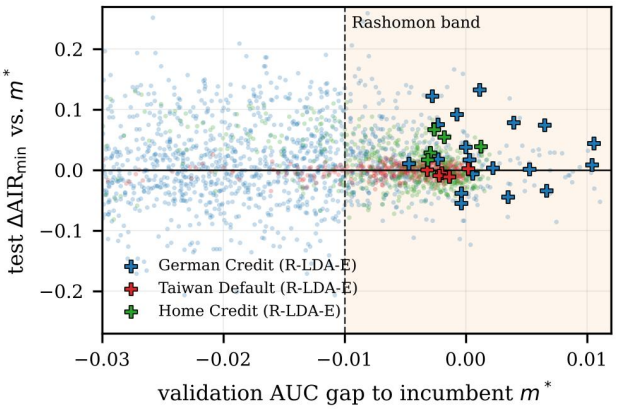

Figure 3 Disparity Multiplicity (S3)

Note: Each dot is a candidate model in one split, plotted relative to the incumbent m^* ; the shaded band is the Rashomon set (ε = 0.01). Crosses mark R-LDA-E for each split.

Table 4 Held-Out Effect of the Selected Alternative versus the Incumbent (Tolerance 0.01)

Dataset	Variant	ΔAIR_{min}	ΔAUC	ΔPDR	Win	Apparent	Realised
German	R-LDA	+0.028	−0.0035	+0.0050	60%	+0.067	+0.028
	R-LDA-E	+0.026	+0.0006	+0.0030	70%	+0.038	+0.026
Taiwan	R-LDA	+0.004	−0.0048	+0.0003	60%	+0.016	+0.004
	R-LDA-E	−0.005	−0.0016	+0.0008	40%	+0.011	−0.005
Home Credit	R-LDA	+0.062	−0.0084	+0.0014	100%	+0.070	+0.062
	R-LDA-E	+0.041	−0.0011	+0.0001	100%	+0.030	+0.041

Note: Win: share of splits with a higher test AIR_{min} than m^* . Apparent/realised: ΔAIR_{min} on validation/test data.

6.3 Comparison with Mitigation Methods

Table 5 contrasts DIRA with the five alternatives. Group thresholds reach the highest AIR on the two larger datasets (0.977 and 0.967) because they equalise approval rates by construction, but they are tier T2: the protected attribute dictates the decision rule, which is disparate treatment under ECOA and the kind of differential treatment that PIPL Article 24 restricts. On German Credit their cut-offs, estimated from small validation groups, do not generalise (AIR 0.863). FairGBM is the strongest tier-T1 method on German Credit and Taiwan Default (AIR 0.909 and 0.939), but it costs 0.024 AUC on German Credit and gains little on Home Credit (0.622), where its error-rate constraint is not tied to the 80% approval policy. Reweighing performs consistently (0.892, 0.919, 0.668) at a small AUC cost, yet it also uses the protected attribute in training. Proxy removal, the input-scrutiny remedy, is the least efficient option: it lowers AUC by 0.084, 0.035 and 0.061 while improving AIR less than R-LDA on German Credit and Home Credit, in line with the input fallacy [17]. Among the tier-T0 methods, which confine the protected attribute to the audit, R-LDA and R-LDA-E offer the best

balance: on Home Credit R-LDA matches Reweighing (0.669 versus 0.668) without using the protected attribute in training, and R-LDA-E does so at one third of Reweighing's AUC cost (−0.0011 versus −0.0034).

Table 5 Worst-Case Test AIR and AUC Change by Method (Mean over Splits)

		German		Taiwan		Home Credit	
Method	Tier	AIR_{min}	ΔAUC	AIR_{min}	ΔAUC	AIR_{min}	ΔAUC
Incumbent m^*	–	0.861	–	0.909	–	0.606	–
Aware	T2	0.823	−0.001	0.913	0.000	0.610	+0.001
Reweighing [21]	T1	0.892	−0.005	0.919	0.000	0.668	−0.003
FairGBM [19]	T1	0.909	−0.024	0.939	−0.003	0.622	0.000
Proxy removal	T0	0.883	−0.084	0.915	−0.035	0.638	−0.061
Group thresh. [20]	T2	0.863	0.000	0.977	0.000	0.967	0.000
R-LDA (ours)	T0	0.889	−0.004	0.913	−0.005	0.669	−0.008
R-LDA-E (ours)	T0	0.887	+0.001	0.905	−0.002	0.647	−0.001

Note: Tier: where the protected attribute enters (T0 audit only, T1 training, T2 decision time).

6.4 Sensitivity to the Tolerance and the Approval Rate

Figure 4 traces the accuracy–fairness frontier as ε grows. On Home Credit the worst-case AIR of R-LDA-E rises from 0.600 at ε = 0 to 0.676 at ε = 0.03, while its AUC loss stays below 0.006; at ε = 0.01 it already secures +0.041 for −0.0011 AUC. On German Credit the ensemble's AUC never falls below the incumbent's for any ε. The approval rate matters as much as the tolerance. On Taiwan Default the incumbent fails the four-fifths screen at α = 0.6 (AIR 0.745), and R-LDA-E lifts it to 0.799 (R-LDA: 0.806), whereas no material alternative exists at α = 0.8. On Home Credit the gains persist across all approval rates (e.g., 0.372 to 0.406 at α = 0.6). The availability of an LDA is thus a property of the deployed policy, not of the data alone.

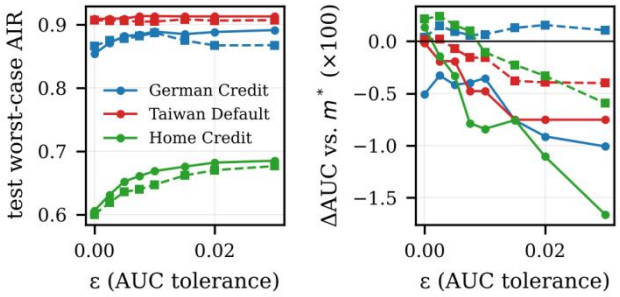

Figure 4 Sensitivity to the Equal-Effectiveness Tolerance ε at α = 0.8
Note: Left: test worst-case AIR; right: AUC change versus m^* . Solid: R-LDA; dashed: R-LDA-E.

6.5 Adverse-Action Multiplicity

Table 6 Adverse-Action Multiplicity within the Rashomon Set

Dataset	Ambig. (%)	Age <25 / ≥25 (%)	Female / male (%)	J	Top-1 flip (%)
German	22.3	29.3 / 21.1	23.0 / 22.0	0.796	57.0
Taiwan	7.4	9.0 / 7.2	7.3 / 7.5	0.765	43.0
Home Credit	14.5	26.1 / 14.1	13.4 / 16.7	0.783	23.7

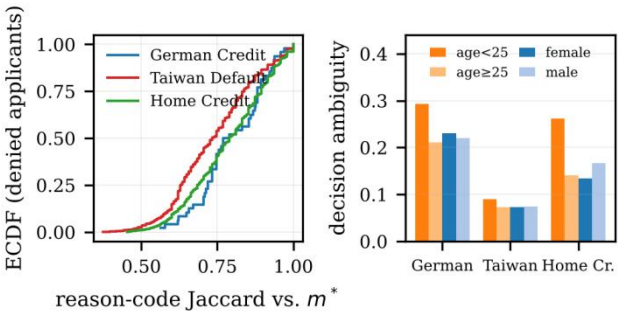

Figure 5 Adverse-Action Multiplicity (S4)
Note: Left: distribution of reason-code agreement J_i for denied applicants (first split). Right: decision ambiguity by group, averaged over splits.

Table 6 and Figure 5 quantify the arbitrariness that remains among equally accurate models. Between 7.4% (Taiwan Default) and 22.3% (German Credit) of test applicants receive conflicting decisions from members of the Rashomon set. The burden falls disproportionately on young applicants: their ambiguity is 26.1% versus 14.1% on Home Credit and 29.3% versus 21.1% on German Credit, and it exceeds that of older applicants in 100%, 85% and 100% of splits (Home

Credit, German Credit, Taiwan Default), echoing the observation that multiplicity weighs more heavily on minority groups [16]. Reasons are unstable as well. The mean Jaccard agreement between the four principal reasons of the incumbent and those of other members is 0.80, 0.76 and 0.78, and for 57.0%, 43.0% and 23.7% of denied applicants at least one equally accurate model reports a different principal reason. The most frequent principal reasons—checking-account status, the latest repayment status and an external bureau score—are themselves among the strongest age proxies in Figure 2, which ties explanation duties directly to proxy discrimination.

7 DISCUSSION: LEGAL AND ETHICAL IMPLICATIONS

7.1 Business Necessity under Multiplicity

The width of the AIR range inside the ε-band shows that the scorecard a lender defends is one draw from a large set of equally effective models. A business-necessity justification that ignores this set is incomplete: the relevant question is whether the disparity of *the chosen* model is necessary, not whether *some* model is predictive. A Rashomon report, i.e., the distribution of accuracy and disparity across the band, gives supervisors and courts a concrete object to examine and fits the bias-examination duty that the AI Act proposal attaches to high-risk credit scoring [7].

7.2 Evidentiary Standards for LDA Claims

Because apparent gains exceeded realised gains by a factor of up to 4.1, an LDA claim, whether advanced by a plaintiff, a regulator or a lender documenting compliance, should be supported by held-out confirmation and uncertainty intervals. Ensembles such as R-LDA-E are a natural safeguard because they turn many weak, noisy alternatives into one stable alternative. Conversely, where repayment behaviour drives the disparity, as on Taiwan Default at $\alpha = 0.8$, the absence of a confirmed alternative strengthens the business-necessity defence, but only for that approval policy.

7.3 Where the Protected Attribute May Enter

The largest reductions in disparity came from methods that use the protected attribute in training (T1) or at decision time (T2). T2 methods conflict with disparate-treatment prohibitions under ECOA and with the ban on unreasonable differential treatment in PIPL Article 24 [3, 11], and the status of T1 methods remains contested [12-13]. DIRA shows that a substantial part of the attainable improvement, on Home Credit the full improvement of Reweighing, can be obtained with the protected attribute confined to the audit. This is consistent with outcome-focused testing and with the AI Act proposal's allowance for processing sensitive data for bias monitoring [7, 17]. Input scrutiny alone is costly and weak; proxy attenuation offers a graded alternative that retains residual predictive value.

7.4 Explanations That Survive Model Choice

Top-1 flip rates of 24–57% mean that the principal reason in an adverse-action notice often reflects model choice rather than the applicant's file. This bears directly on the specificity requirement of Regulation B and the CFPB circulars [3, 5-6], on the explanation rights confirmed in *SCHUFA* and in Article 18 of the Consumer Credit Directive [9-10], and on PIPL Article 24 [11]. Lenders should (i) report reason stability alongside accuracy, (ii) disclose principal reasons that are shared across the Rashomon set, and (iii) route applicants with unstable reasons or ambiguous decisions to human review, which the Consumer Credit Directive already guarantees on request.

7.5 Ethical Implications

Arbitrariness is not evenly distributed. Young applicants, who already face lower approval rates, also face the highest probability that their decision hinges on an arbitrary modelling choice, so a procedural harm compounds the distributive one [15-16]. Ensembling and Rashomon documentation are low-cost steps toward procedural fairness. Because DIRA holds the approval volume fixed and reports the portfolio default rate for every alternative, inclusion gains are never purchased with irresponsible lending. The audit operates on booked loans with observed protected attributes; in deployment its audit stage can run on estimated group membership, and the candidate pool can be extended to other model families without changing the procedure.

8 CONCLUSION

This paper translated the burden-shifting logic of fair-lending law and the duty to explain adverse decisions into DIRA, a four-stage audit for AI-driven lending. On three public credit datasets, proxies made protected attributes highly recoverable, equally accurate models differed widely in disparity, and a Rashomon ensemble that never sees protected attributes in training or prediction found less discriminatory alternatives at negligible accuracy cost wherever such alternatives existed. The study also showed that in-sample LDA claims overstate attainable gains and that both decisions and adverse-action reasons vary across equally accurate models, with the variation concentrated on young applicants. Three policy positions follow: disparity should be assessed against the Rashomon set rather than a single model; LDA

claims require held-out confirmation; and reason stability should become part of explainability compliance. Future work will extend DIRA to pricing decisions, intersectional groups and settings in which protected attributes are inferred.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Fuster A, Goldsmith-Pinkham P, Ramadorai T, et al. Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 2022, 77(1): 5-47.
[2] Bartlett R, Morse A, Stanton R, et al. Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 2022, 143(1): 30-56.
[3] Equal Credit Opportunity Act, 15 U.S.C. §1691 et seq.; Regulation B, 12 C.F.R. pt. 1002, 2023.
[4] Texas Dept. of Housing and Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. 519, 2015.
[5] Consumer Financial Protection Bureau. Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms. Consumer Financial Protection Bureau, 2022.
[6] Consumer Financial Protection Bureau. Consumer Financial Protection Circular 2023-03: Adverse action notification requirements and the proper use of the CFPB's sample forms provided in Regulation B. Consumer Financial Protection Bureau, 2023.
[7] European Commission. Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final, Brussels, European Commission, 2021.
[8] Regulation (EU) 2016/679 (General Data Protection Regulation), Official Journal of the European Union, L 119, , 2016: 1-88.
[9] Court of Justice of the European Union, Case C-634/21, OQ v Land Hessen (SCHUFA Holding), Judgment of 7 Dec. 2023, ECLI:EU:C:2023:957.
[10] Directive (EU) 2023/2225 of the European Parliament and of the Council of 18 October 2023 on credit agreements for consumers and repealing Directive 2008/48/EC, Official Journal of the European Union, 2023.
[11] Personal Information Protection Law of the People's Republic of China, adopted Aug. 20, 2021, in force Nov. 1, 2021, Art. 24.
[12] Kumar I E, Hines K E, Dickerson J P. Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with U.S. fair lending regulation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2022. DOI: 10.1145/3514094.3534154.
[13] Wachter S, Mittelstadt B, Russell C. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review, 2021, 41: 105567.
[14] Marx C T, du Pin Calmon F, Ustun B. Predictive multiplicity in classification. In: Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, 2020, 119: 6765-6774.
[15] Black E, Raghavan M, Barocas S. Model multiplicity: Opportunities, concerns, and solutions. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2022: 850-863.
[16] Watson-Daniels J, Parkes D C, Ustun B. Predictive multiplicity in probabilistic classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(9): 10306-10314.
[17] Gillis T B. The input fallacy. Minnesota Law Review, 2022, 106(3): 1175.
[18] Prince A E R, Schwarcz D. Proxy discrimination in the age of artificial intelligence and big data. Iowa Law Review, 2020, 105(3): 1257-1318.
[19] Cruz A F, Belém C, Jesus S, et al. FairGBM: Gradient boosting with fairness constraints. In: Proceedings of the International Conference on Learning Representations (ICLR), 2023.
[20] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, 2016, 29: 3315-3323.
[21] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 2012, 33(1): 1-33.
[22] Kozodoi N, Jacob J, Lessmann S. Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 2022, 297(3): 1083-1094.
[23] Barocas S, Selbst A D. Big data's disparate impact. California Law Review, 2016, 104(3): 671-732.
[24] Hall P, Cox B, Dickerson S, et al. A United States fair lending perspective on machine learning. Frontiers in Artificial Intelligence, 2021, 4: 695301.
[25] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2(1): 56-67.
[26] Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §1607.4(D) (1978).
[27] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, 2017, 30: 3146-3154.
[28] Hofmann H. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: 10.24432/C5NC77.

[29] 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 Systems with Applications, 2009, 36(2): 2473-2480.
[30] Home Credit Group. Home Credit Default Risk. Kaggle competition, 2018. Available: https://www.kaggle.com/c/home-credit-default-risk.