Fairness by Design: Machine Learning and Interpretable Mortgage Lending*

Sebastian Bell KIT

Ali Kakhbod UC Berkeley

Abdolreza Nazemi KIT

Richard Stanton UC Berkeley

Nancy Wallace UC Berkeley

Abstract

Regulated lenders must explain denials and avoid disparate treatment in U.S. mortgages, yet flexible scoring can be opaque. We develop and study a less discriminatory and transparent machine-learning approach that embeds equalized-odds fairness in estimation and decomposes each decision into feature contributions suitable for adverse-action notices. Using HMDA applications, the model preserves performance while shrinking minority–nonminority error gaps and reweighting away from geographic proxies toward core underwriting signals (debt-to-income, loan-to-value). Supplementary analyses around underwriting cutoffs and lender-level tests indicate that the model lowers minority shortfalls at the margin and reduces within-lender disparities.

JEL Classification: C45, G21, G28, R30

Keywords: Credit, Mortgages, Machine learning, Fairness, Interpretability, Regulations.

*First Draft: November 2025. This draft: February 2026. Sebastian Bell is with the School of Economics and Management, Karlsruhe Institute of Technology (KIT), email: sebastian.bell@kit.edu. Ali Kakhbod is with the Haas School of Business at UC Berkeley, Berkeley, CA 94720, email: akakhbod@berkeley.edu. Abdolreza Nazemi is with the School of Economics and Management, Karlsruhe Institute of Technology (KIT), email: nazemi@kit.edu. Richard Stanton is with the Haas School of Business at UC Berkeley, Berkeley, CA 94720, email: richard.stanton@berkeley.edu. Nancy Wallace is with the Haas School of Business at UC Berkeley, Berkeley, CA 94720, email: newallace@berkeley.edu.

Contents

1 Introduction **1**

2 Estimation methodology **6**
2.1 Fairness objective . . . 7
2.2 Theoretical properties of the fairness penalty . . . 9
2.3 Training algorithm . . . 14

3 Data and Results **14**
3.1 Data . . . 15
3.2 Performance evaluation . . . 16
3.3 Predictor importance . . . 17
3.4 Shape functions . . . 21
3.5 Interactions . . . 22

4 Behavior at underwriting thresholds **26**

5 Robustness and additional analysis **27**
5.1 Adverse action reasons . . . 30
5.2 Decomposition of the EO disparity . . . 32
5.3 Within-lender fairness . . . 37
5.4 Risk parity and pricing alignment . . . 39

6 Conclusions **42**

A Summary statistics **44**

B Variable definitions **47**

C Model: lender profit maximization **50**

D Additional algorithmic details **54**
D.1 Preprocessing and parameterization . . . 54
D.2 Warm starts . . . 55
D.3 Training objective and regularization . . . 56
D.4 Interaction discovery . . . 58
D.5 Pruning . . . 59
D.6 Three-stage optimization and early stopping . . . 60
D.7 Inference details . . . 61

E Omitted proofs **62**
E.1 Proof of Proposition 1 . . . 63
E.2 Proof of Proposition 2 . . . 64
E.3 Proof of Proposition 3 . . . 65

F Empirical evaluation of penalty **75**

G Robustness: Threshold variation **77**

H Robustness: Alternative cutoffs **81**
H.1 Threshold analysis using DTI cutoff 81
H.2 Placebo test 84
H.3 Uniform kernel estimation 86

I Robustness: Removing other sensitive attributes **89**

J Robustness: Alternative data **90**

K Robustness: Bank-lender jumbo loan subsample **94**

L Robustness: Subgroup fairness **95**

M Robustness: Decomposition of the global score gap **98**

1 Introduction

Machine learning (ML) models are increasingly used to evaluate credit applications, promising improved risk prediction and operational efficiency; in consumer lending, however, these gains face binding legal and institutional constraints. U.S. fair-lending and model-risk regimes—including the Equal Credit Opportunity Act (ECOA) and its implementing Regulation B; the Fair Credit Reporting Act (FCRA); and supervisory guidance under SR 11-7—require that credit decisions be transparent and well documented, and that creditors provide applicants with specific reasons for adverse action (Board of Governors of the Federal Reserve System, 2011; Consumer Financial Protection Bureau, 2022). Regulation B requires that adverse-action notices "be specific and indicate the principal reason(s) for the adverse action" and that those reasons "relate to and accurately describe the factors actually considered or scored by a creditor" (12 C.F.R. §1002.9(b)(2)). Complementing this requirement, SR 11-7 emphasizes that models used in high-stakes decisions must be "well understood by users," supported by "clear documentation of model design, limitations, and assumptions," and capable of effective challenge and explanation (Board of Governors of the Federal Reserve System, 2011). Consistent with these principles, recent CFPB guidance cautions that when creditors rely on complex or opaque models, they remain obligated to provide adverse-action reasons that are accurate and specific to the individual decision, and that models whose internal logic cannot be meaningfully explained may be unable to satisfy this requirement in practice (Consumer Financial Protection Bureau, 2022).

These regulatory requirements have led regulators, courts, and practitioners to emphasize model transparency and to warn against opaque "black-box" decision systems. The Federal Reserve has highlighted the risks posed by model opacity in financial services (Brainard, 2021), while courts have emphasized that adverse-action disclosures serve a core antidiscrimination function by requiring lenders to articulate the actual basis for individual credit denials (Treadway v. Gateway Chevrolet Oldsmobile Inc., 2004). A common industry response is therefore to rely on simple, linear scorecards, which are often viewed as "easy to explain and use" (Ustun and Rudin, 2019). This response implicitly treats regulatory explainability as incompatible with modern, nonlinear prediction.

At the same time, both legal and technical literatures caution that transparency cannot be reduced to mechanical simplicity alone. In regulated credit settings, fairness is inherently relational: outcomes must be assessed relative to feasible alternative decision rules, not against abstract mathematical benchmarks (Lee and Floridi, 2021). Complementing this perspective, work in interpretable machine learning emphasizes that post-hoc ex-

planations of complex models may be unstable or misleading, and instead advocates for models that are transparent by construction, with internal structure that supports faithful, instance-level explanations (Alvarez-Melis and Jaakkola, 2018; Rudin, 2019; Sudjianto and Zhang, 2021). Together, these insights suggest that neither linear scorecards nor opaque machine-learning systems provide a satisfactory foundation for making compliance-relevant credit decisions.

Recent evidence also shows that neither the status quo nor unconstrained algorithmic sophistication reliably improves equity in mortgage markets. Using institutional features of GSE and FHA pricing grids, Bartlett et al. (2022b) show that risk-equivalent Black and Hispanic borrowers pay systematically higher rates than otherwise similar non-minority borrowers, consistent with impermissible discrimination. Complementing these findings, Fuster et al. (2022) show that more flexible ML methods can increase dispersion in predicted risk and worsen within-group outcomes for minority borrowers, even as average acceptance rises. Together, these results suggest that both linear scorecards and unconstrained ML can fail to address—and may even exacerbate—disparities.

We complement the important work by Fuster et al. (2022) by studying the trade-off between flexibility, fairness, and interpretability. The key innovation is that we integrate three elements that the existing credit market literature usually treats separately: inherent interpretability, fairness (minimizing disparities across groups), and maximizing predictive accuracy. Specifically, we construct an alternative model of historical underwriting outcomes that can be interpreted as a less discriminatory alternative (LDA) within a common public-information set: an algorithm that simultaneously (i) reduces group disparities (e.g., minority vs. non-minority) in error rates relative to observed denial/origination labels, (ii) preserves strong out-of-sample fit to historical underwriting labels, and (iii) remains transparent enough to satisfy compliance and adverse-action requirements. Using U.S. mortgage data, we show that such an LDA can be implemented in practice and delivers sizable reductions in label-conditional disparity with minimal loss in predictive accuracy.

In particular, we develop the *Fairness-aware Additive Network* (FAN), a classifier designed to combine fairness-constrained estimation with an inherently interpretable structure. To showcase the methodology, we apply FAN to mortgage outcomes (rejection versus origination) in the public-HMDA information set. As such, FAN allows us to bring fairness-aware estimation and inherent interpretability together in a single modeling framework suited to regulated credit decisions.

Our contribution is threefold. First, we introduce an in-processing EO regularization that directly shapes the underwriting decision boundary during estimation. Theoretically,

we show that, even though the penalty is threshold-free and differentiable, it is decision-focused. It concentrates weight near the approval cutoff where decisions flip and, therefore, targets the economically relevant margin. Furthermore, we show that the penalty term is bounded by a disparity measure across decision thresholds, giving the penalty a transparent interpretation in terms of decision-relevant cross-group differences in model behavior. Finally, we establish that, under our architectural constraints, the penalty generalizes out of sample, so the fairness improvements are not merely in-sample artifacts.

Second, we demonstrate that this fairness constraint can be embedded in a neural additive architecture that retains substantial nonlinear predictive power while remaining interpretable by construction. Because the model is additive at the feature level, each prediction can be decomposed exactly into variable-specific contributions, making it possible to document how the decision rule changes under fairness regularization and to map those changes into adverse-action-style explanations without relying on potentially unstable post-hoc explanation tools.

Third, we show that meaningful fairness gains can be achieved even within a public, harmonized application-information set by applying FAN to public HMDA approval and denial decisions. This implies that, within the public-information environment we study, less disparate alternatives can be constructed without relying on proprietary underwriting inputs or opaque vendor models. We show empirically that the EO-regularized training materially reduces minority–non-minority disparities in model-implied mortgage denials with negligible loss in out-of-sample predictive fit to observed underwriting labels. Exploiting FAN's separable structure, we show how fairness adjustments operate through economically meaningful channels. EO regularization shifts weight away from proxy-like variables correlated with protected status but less central for fitting observed denial/origination outcomes, and toward core underwriting signals such as debt-to-income and loan-to-value ratios. Using local regression-discontinuity designs around underwriting thresholds, we find that minority shortfalls at the decision margin decline as the EO penalty increases. Importantly, these gains do not disrupt risk–price relationships: conditional on predicted risk, approved minority and non-minority borrowers face similar interest rates.

Taken together, our results show that compliance-relevant explainability does not require retreating to purely linear models, and that increased algorithmic sophistication need not come at the expense of producing less disparate model decisions. When fairness and interpretability are treated as first-order design objectives, ML can reduce model-implied disparities while preserving predictive performance and regulatory transparency.

Relation to the literature

This paper contributes to four strands of literature by linking recent advances in machine learning, empirical evidence on mortgage discrimination, and legal doctrines of disparate impact to a practical, fairness-aware, and interpretable modeling framework.

Machine learning in finance and its impact on credit outcomes. A growing literature applies machine learning methods to financial prediction, including neural networks (e.g., Chen et al., 2024a; Fallahgoul et al., 2024; Gu et al., 2021), tree-based methods (e.g., Bell et al., 2025; Bryzgalova et al., 2025; Lopez-Lira and Roussanov, 2023; van Binsbergen et al., 2023), and generative models (e.g., Chen et al., 2024b; Kakhbod et al., 2025). In mortgage markets, flexible ML models have been used to forecast default risk and study distributional effects in pricing and approvals (Fuster et al., 2022; Jansen et al., 2025). A central finding in this literature is that increased model flexibility often raises dispersion in predicted risk and, in equilibrium, widens within-group spreads in prices and approvals—even when average access improves (Fuster et al., 2022). Related evidence from automated credit scoring highlights both accuracy gains and the risk that high-dimensional models exacerbate racial disparities through correlated proxies (Berg et al., 2020; Howell et al., 2024; O'Neil, 2016).

Our contribution builds on this work by explicitly governing these distributional effects during estimation. Rather than relying on unconstrained flexibility or post-hoc adjustments, we combine nonlinear predictive capacity with an in-processing fairness penalty and structural interpretability, allowing lenders and regulators to directly observe how fairness constraints reshape the learned decision rule.

Racial and ethnic disparities in mortgage markets. A large empirical literature documents persistent racial and ethnic disparities in mortgage pricing and outcomes.[1] Exploiting institutional details of the GSE and FHA pricing grids, Bartlett et al. (2022b) show that Black and Latinx borrowers are charged higher interest rates than risk-equivalent non-minority borrowers, consistent with prohibited disparate treatment. Related work shows that high-cost lending, lender market structure, and targeted advertising disproportionately affect minority communities (Bayer et al., 2018; Gurun et al., 2016).

These findings motivate approaches that make disparities traceable to specific risk factors and decision rules. Our additive, separable architecture directly serves this goal

[1] This is part of a long literature finding minority/non-minority differences in both mortgage-approval probabilities and interest rates, including Ambrose et al. (2021); Bayer et al. (2018); Bhutta and Hizmo (2021); Black et al. (1978, 2003); Cheng et al. (2015); Courchane and Nickerson (1997); Ghent et al. (2014); Hanson et al. (2016); Munnell et al. (1996); Reid et al. (2017) and Willen and Zhang (2025). Similar findings in other markets include Dobbie et al. (2021) and Butler et al. (2023).

by decomposing denials and approvals into economically interpretable components, facilitating both diagnosis of disparities and compliance with adverse-action requirements.

Law and economics of algorithmic fairness and discrimination. The law and economics literature has shifted from asking whether algorithms can be fair to how fairness can be made legally and empirically demonstrable. Even when protected characteristics are excluded, correlated proxies can reintroduce discrimination, rendering purely formal compliance insufficient (Gillis and Spiess, 2019). As a result, scholars emphasize ex ante testing, auditing, and comparison of alternative decision rules across lenders and vendors (Kim, 2017; Kroll et al., 2017). Complementing these accountability paradigms, Kleinberg et al. (2020) argue that algorithms can also act as "Geiger counters" for bias by detecting disparate treatment with precision.

A related strand moves from documenting outcome disparities to clarifying which disparities are informative for assessing discrimination under legal standards. Yang and Dobbie (2020) and Bartlett et al. (2022a) focus on causal input fairness, which better aligns machine-learning design with evidentiary standards for disparate impact. Recent work also highlights the legal significance of model multiplicity: Black et al. (2024) find model multiplicity nearly always exists; thus under the logic of disparate impact doctrine, developers have a duty to search for "less discriminatory alternatives" (LDAs). Normatively, Hellman (2020) argues that predictive parity pertains to belief whereas fairness concerns action, making disparities in error rates a prima facie indicator of injustice.

Methodologically, the computer science literature formalizes these ideas through group fairness criteria such as equalized odds (Hardt et al., 2016), while demonstrating the incompatibility of multiple desiderata such as calibration and error-rate parity (Kleinberg et al., 2017). Most applied work evaluates fairness post hoc and mitigates disparities through thresholding or feature neutralization (e.g., Hurlin et al., 2025).

Our approach differs by embedding a differentiable equalized-odds penalty directly into model estimation. This ensures that fairness considerations shape the learned decision rule itself—precisely the object of legal scrutiny—while pairing this objective with inherent interpretability rather than post-hoc explanation.

Borrower and market frictions in underwriting and pricing. A related literature studies how behavioral and institutional frictions shape credit outcomes. Household mistakes in take-up and refinancing generate persistent rate dispersion (Agarwal et al., 2017; Keys et al., 2016), while program design and guarantees modulate how capacity and collateral risks are screened (Gerardi et al., 2023).

Our model complements this work by recovering nonlinear shape functions for core

underwriting variables—such as debt-to-income, loan-to-value, and loan purpose—and by showing how fairness regularization reallocates decision weight away from proxy-like variables toward fundamental risk channels.

Overall, while prior literature documents disparities and the distributional consequences of machine learning in credit markets, we provide a practical framework for training fairness-aware models that remain economically interpretable and legally meaningful. By linking evidence on mortgage discrimination (Bartlett et al., 2022b; Bayer et al., 2018; Bhutta and Hizmo, 2021) and ML-induced dispersion (Berg et al., 2020; Fuster et al., 2022; Howell et al., 2024) to an in-processing, interpretable design, our approach operationalizes the concept of less discriminatory alternatives in a form suitable for regulated lending.

The rest of this paper is organized as follows. Section 2 details the estimation methodology. Section 3 presents our data set and empirical results. Sections 4 and 5 provide extensions and robustness tests. Section 6 concludes.

2 Estimation methodology

To construct a *less discriminatory alternative* (LDA) that integrates fairness directly into estimation while remaining transparent enough for regulatory use, we develop a customized machine learning model: the *Fairness-aware Additive Network* (FAN).

FAN is a fairness-aware additive model that combines the interpretability of generalized additive model (GAM)-style decompositions with the flexibility of modern machine learning,[2] while allowing fairness objectives to shape the learned decision rule itself. The model assumes a separable structure in which predicted outcomes are decomposed into main effects and a sparse set of pairwise interactions, facilitating both global interpretability and exact instance-level attribution.

Let $l = 1, \ldots, L$ index loan applications in our dataset. For each application l, we denote the feature vector by $X_l = (x_i)_l \in \mathbb{R}^N$ (with N predictors indexed by i) and the observed outcome by $y_l \in \{0, 1\}$ as defined above. The FAN model is parameterized by θ

[2]For general introductions to the family of GAMs, see Hastie and Tibshirani (1986), Yang et al. (2021) and Stasinopoulos et al. (2024). Our departure from this literature is to incorporate fairness constraints directly into the estimation problem and study their empirical and theoretical properties, yielding models that are fairness-aware and transparent by construction while allowing for smoothness, sparsity, and (when desired) monotonicity in selected features.

and produces a scalar score

$$s_\theta(X_l) = F_\theta(X_l) = \beta_0 + \sum_{i=1}^{N} f_i(x_{i,l}) + \sum_{(i,j)\in\mathcal{I}} f_{ij}(x_{i,l}, x_{j,l}), \tag{1}$$

where f_i are one-dimensional shape functions capturing the main effect of predictor i, $\mathcal{I}$ is a learned sparse set of interactions, and f_{ij} are two-dimensional shape functions representing pair-wise interactions between inputs. The final model output, the predicted probability of rejection $\hat{p}_l$ for application l, is then computed by applying the sigmoid function to the raw score s_l:

$$\hat{p}_l(\theta) = \sigma\big(s_\theta(X_l)\big), \qquad \text{with } \sigma(u) = \tfrac{1}{1+e^{-u}}.$$

This additive decomposition yields globally interpretable effect curves $f_i(\cdot)$ for each predictor, a small number of interpretable interactions $f_{ij}(\cdot,\cdot)$, and additive attributions that can, for example, be linked to adverse action reasons.

2.1 Fairness objective

In order to fit the model parameters θ, we minimize a loss function originally defined at the level of individual applications. By default, the classifier minimizes standard binary cross-entropy (BCE) loss given by:

$$\ell_l^{\text{BCE}}(\theta) = -\Big[y_l \log \hat{p}_l(\theta) + \big(1 - y_l\big) \log\big(1 - \hat{p}_l(\theta)\big)\Big]$$

and the overall loss is the average of ℓ_l^{BCE} across the training sample. This objective encourages accurate fit to the classification targets but does not by itself impose any constraint on group fairness. The flexibility of our machine learning models allows us to expand the objective in order to also take algorithmic fairness into account.

One challenge in defining an extended loss function is ensuring differentiability with respect to the model parameters, which is necessary to apply gradient descent during training. Equalized odds requires parity of true-positive rates (TPR) and false-positive rates (FPR) across groups at a common decision rule. Hard-thresholding the score to compute exact TPR/FPR would violate differentiability and introduce sensitivity to the particular threshold chosen. Instead, we construct a *threshold-free, score-based proxy* that aligns the class-conditional score distributions across groups.

Consider a subsample of applications used in one training step, indexed by l =

$1, \ldots, B$.[3] Let $a_l \in \mathcal{G}$ denote the group membership (e.g., racial or ethnic group) associated with an individual loan application l, and define the indicators $m_l^+ = \mathbb{1}\{y_l = 1\}$ for rejected loans and $m_l^- = \mathbb{1}\{y_l = 0\}$ for originated loans. For each group $g \in \mathcal{G}$, let

$$D_g^+ = \sum_{l=1}^{B} m_l^+ \, \mathbb{1}\{a_l = g\}, \qquad D_g^- = \sum_{l=1}^{B} m_l^- \, \mathbb{1}\{a_l = g\},$$

be the number of rejected and originated applications for group g in the subsample.

We then define soft estimates of group-specific TPR and FPR as averages of the predicted scores:

$$\widehat{\mathrm{TPR}}_g(\theta) = \frac{\sum_{l=1}^{B} \hat{p}_l(\theta) \, m_l^+ \, \mathbb{1}\{a_l = g\}}{D_g^+ + \varepsilon}, \qquad \widehat{\mathrm{FPR}}_g(\theta) = \frac{\sum_{l=1}^{B} \hat{p}_l(\theta) \, m_l^- \, \mathbb{1}\{a_l = g\}}{D_g^- + \varepsilon},$$

with $\varepsilon > 0$ included for numerical stability. Intuitively, these quantities are the average predicted rejection probabilities for actually originated and actually rejected applications in each group. If a group has no originated or no rejected loans in the subsample, the corresponding term is omitted.

To encourage parity across groups, we penalize the dispersion of these estimates. Specifically, let $\overline{\mathrm{TPR}}$ and $\overline{\mathrm{FPR}}$ denote the mean of the group-specific estimates across groups with valid denominators, i.e. $\overline{\mathrm{TPR}}(\theta) = \frac{1}{|\mathcal{G}_+|} \sum_{g \in \mathcal{G}_+} \widehat{\mathrm{TPR}}_g(\theta)$ and $\overline{\mathrm{FPR}}(\theta) = \frac{1}{|\mathcal{G}_-|} \sum_{g \in \mathcal{G}_-} \widehat{\mathrm{FPR}}_g(\theta)$. The fairness penalty is defined as

$$\mathcal{P}_{\mathrm{EO}}(\theta) = \frac{1}{|\mathcal{G}_+|} \sum_{g \in \mathcal{G}_+} \left(\widehat{\mathrm{TPR}}_g(\theta) - \overline{\mathrm{TPR}}(\theta)\right)^2 + \frac{1}{|\mathcal{G}_-|} \sum_{g \in \mathcal{G}_-} \left(\widehat{\mathrm{FPR}}_g(\theta) - \overline{\mathrm{FPR}}(\theta)\right)^2,$$

where $\mathcal{G}_+$ and $\mathcal{G}_-$ denote the set of groups with at least one positive and at least one negative observation, respectively. This term measures the variance of group-level TPRs and FPRs around their averages, and is minimized when the group-specific rates coincide.[4]

[3] Such a subsample is usually called a "(mini-)batch" in backpropagation algorithms.

[4] In our application, we consider two groups, i.e., the protected attribute a_l is binary: $a_l = 1$ for minority applicants and $a_l = 0$ for non-minority applicants. When both groups are present in the minibatch, $\mathcal{G}_+ = \mathcal{G}_- = \{0, 1\}$ and $\overline{\mathrm{TPR}}(\theta) = \frac{1}{2}\left(\widehat{\mathrm{TPR}}_0(\theta) + \widehat{\mathrm{TPR}}_1(\theta)\right)$ (and analogously for $\overline{\mathrm{FPR}}(\theta)$). Substituting into the general penalty gives $\frac{1}{2}\sum_{g=0}^{1}(\widehat{\mathrm{TPR}}_g - \overline{\mathrm{TPR}})^2 = \frac{1}{4}(\widehat{\mathrm{TPR}}_1 - \widehat{\mathrm{TPR}}_0)^2$, and the same identity holds for FPR. Thus, up to the constant factor $\frac{1}{4}$ (absorbed into λ_{EO}), the fairness penalty reduces to a simple squared difference between the two groups' soft rates:

$$\mathcal{P}_{\mathrm{EO}}(\theta) = \left(\widehat{\mathrm{TPR}}_1(\theta) - \widehat{\mathrm{TPR}}_0(\theta)\right)^2 + \left(\widehat{\mathrm{FPR}}_1(\theta) - \widehat{\mathrm{FPR}}_0(\theta)\right)^2.$$

Finally, the overall loss function combines predictive accuracy and fairness. For each loan application l, we define

$$\ell_l(\theta) = \ell_l^{\text{BCE}}(\theta) + \lambda_{\text{EO}} \mathcal{P}_{\text{EO}}(\theta),$$

where $\lambda_{\text{EO}} \geq 0$ is a hyperparameter weighting the fairness penalty. The loss used for backpropagation updates of the parameters is the mean of $\ell_l(\theta)$ over the subsample (batch) of size B:

$$\mathcal{L}_{\text{subsample}}(\theta) = \frac{1}{B} \sum_{l=1}^{B} \ell_l(\theta).$$

This structure of the loss function can be economically motivated by modeling lenders as risk-neutral profit maximizers, who approve a loan whenever the expected return exceeds the expected loss under a regulatory (or self-imposed) fairness constraint.[5] By increasing λ_{EO}, the training process places more emphasis on aligning group-specific true and false positive rates, thereby moving the classifier towards the equalized odds criterion. Note that the penalty term is freely adjustable, so alternative measures of algorithmic fairness can be incorporated as well.

In the empirical application below, equalized odds is defined with respect to the observed HMDA denial/origination label, not realized repayment or default. Accordingly, the penalty equalizes model behavior conditional on historical underwriting outcomes rather than latent credit risk.

2.1.1 Deployment

It is worth clarifying the distinction between training and deployment. We train the scoring model $s_\theta(\cdot)$ by minimizing $\mathcal{L}_{\text{subsample}}(\theta)$, a differentiable objective consisting of the prediction loss and an equalized-odds *surrogate* based on *soft* group rates—i.e., within-cell averages of $\hat{p}_l(\theta) = \sigma\big(s_\theta(X_l)\big)$—and therefore the optimization does not involve any hard threshold. This yields the estimated model parameters θ from the training data. In practice, at deployment we convert scores to binary decisions using a hard cutoff, taking $\widehat{y} = \mathbf{1}\{s_\theta(X) \geq \tau\}$; for example, we set $\tau = 0$ (equivalently $\hat{p}_\theta(X) \geq 0.5$).

2.2 Theoretical properties of the fairness penalty

So far, we have defined the model's fairness penalty as the variance across groups in the average predicted rejection probability, summed over both actually rejected and actually

[5] A conceptual formalization is provided in Appendix C.

approved loan applications. This fairness term adds to the BCE a scalar penalty that, for each label $y \in \{0,1\}$, measures the dispersion across groups $g \in \mathcal{G}$ of the conditional mean predicted probability, $\mu_{g,y}(\theta) := \mathbb{E}[\sigma(s_\theta(X)) \mid A{=}g, Y{=}y]$, where $\mu_{g,1}$ is the average denial score the model assigns to truly positive cases ($Y{=}1$) within group g and $\mu_{g,0}$ is the average denial score for truly negative cases ($Y{=}0$). Equalizing these conditional means across groups reduces systematic score differences that would otherwise translate into different error rates in model predictions.

What is still missing is a direct connection between this soft, threshold-agnostic penalty formulation and fairness outcomes that matter when applying the model in practice: In contrast to the training stage, applying the model in practice requires a hard threshold τ that turns the continuous model output into a binary decision. In the following, we relate our penalty term to practically relevant equalized-odds behavior at this score cutoff where decisions shift from approval to rejection. Specifically, we ask: Does this penalty actually act on fairness where approvals are decided? Can we translate it into transparent bounds that summarize fairness at the decision cutoff? And will those guarantees hold up out of sample? In the following section we provide answers to these questions. In particular, we show (i) why the penalty actually focuses on a relevant decision boundary rather than being spread over arbitrary thresholds, (ii) how it relates to interpretable, quantitative bounds connected to group error-rate disparities, and (iii) whether and under what sample sizes and group-label prevalences these guarantees remain reliable beyond the training data.

We do so by providing theoretical justifications for three key properties of the penalty function.[6] First, we demonstrate that even though the penalty does not fix a single threshold (in order to ensure differentiability), it naturally puts most of its attention on scores near the approval cutoff, i.e., where the approve/deny decision flips (Proposition 1). Second, we provide explicit, threshold-agnostic bounds on average EO disparities (Proposition 2), yielding summaries of how unequal error rates are in the neighborhood that drives approvals and denials. Third, we establish that these fairness quantities generalize beyond the training sample $\mathcal{T}_1$ with $n := |\mathcal{T}_1|$ at the usual $1/\sqrt{n}$ rate, scaled by the minimal group-label cell mass $p_{\min}$ (Proposition 3), which allows us to formulate data requirements that ensure dependable deployment. Together, these results make the fairness weight λ a practical control in line with governance objectives: increasing it moves the model toward smaller differences between groups right where decisions are made in practice.

First, the fairness term is threshold-agnostic, but nevertheless targets the region where decisions are made. While avoiding a hard threshold, it matches a weighted average of

[6] Detailed proofs of these properties are provided in Appendix E.

groups' acceptance/denial rate curves, with extra emphasis near the decision boundary, so it targets fairness where choices flip, at the probability threshold of 0.5:

Proposition 1. *For any* (g, y)*,*

$$\mu_{g,y}(\theta) \;=\; \mathbb{E}\big[\sigma(S) \mid A{=}g, Y{=}y\big] \;=\; \int_{-\infty}^{\infty} \sigma'(\tau)\, \Pr\big(S \geq \tau \mid A{=}g, Y{=}y\big)\, d\tau \;=\; \int_{-\infty}^{\infty} w(\tau)\, R_{g,y}(\tau)\, d\tau,$$

where $R_{g,y}(\tau)$ *is the TPR curve when* $y{=}1$ *and the FPR curve when* $y{=}0$*, and* $w(\tau) = \sigma'(\tau) = \sigma(\tau)\big[1 - \sigma(\tau)\big]$ *is a bell-shaped weight that peaks at the decision boundary.*

Conceptually, this identity shows that aligning $\mu_{g,y}$ across groups is equivalent to aligning a smoothed average of their TPR/FPR curves with the smoothing weight concentrated where classifications flip. The penalty therefore encourages the score distributions to overlap in the neighborhood of the relevant threshold without requiring a specific hard cutoff. This has immediate benefits in practice, as (a) fairness is not fragile to small shifts in the operating threshold, because the training signal already emphasized a band of thresholds around the boundary and (b) gradients remain smooth and numerically stable because the surrogate avoids non-differentiable hard decisions. In other words, the proposition expresses that the conditional mean probability equals a weighted average of the group's rate (EO) curve across thresholds τ:

$$\mu_{g,y}(\theta) = \int_{-\infty}^{\infty} \underbrace{\sigma'(\tau)}_{w(\tau)} \underbrace{\Pr\big(s_\theta(X) \geq \tau \mid A{=}g, Y{=}y\big)}_{R_{g,y}(\tau)}\, d\tau,$$

where the weight $w(\tau) = \sigma'(\tau) = \sigma(\tau)[1 - \sigma(\tau)]$ is a bell-shaped density centered at the decision boundary ($\tau = 0$ where $\hat{p} = 0.5$). Therefore, the penalty equalizes across groups the w-weighted averages of the TPR ($y{=}1$) and FPR ($y{=}0$) curves, staying focused on the operational region without enforcing an explicit (undifferentiable) threshold.

Second, the penalty is upper-bounded by a weighted EO disparity across thresholds:

Proposition 2. *For each* $y \in \{0, 1\}$*, denote* $\mathbb{E}_w[R_{g,y}] := \int_{-\infty}^{\infty} w(\tau)\, R_{g,y}(\tau)\, d\tau$*. Then,*

$$\underbrace{\mathrm{Var}_g\big(\mathbb{E}_w[R_{g,y}]\big)}_{\textit{our penalty piece for label } y} \;\leq\; \int_{-\infty}^{\infty} w(\tau)\; \mathrm{Var}_g\big(R_{g,y}(\tau)\big)\, d\tau \;=:\; \mathcal{D}_y(\theta).$$

Consequently,

$$\mathcal{P}_{\text{mean}}(\theta) \;:=\; \mathrm{Var}_{g \in \mathcal{G}}\big(\mu_{g,1}(\theta)\big) \;+\; \mathrm{Var}_{g \in \mathcal{G}}\big(\mu_{g,0}(\theta)\big) \;\leq\; \mathcal{D}_1(\theta) + \mathcal{D}_0(\theta).$$

This property further underlines why the penalty term is economically targeted: The cross-group disparity $\mathrm{Var}_g(\mu_{g,1})+\mathrm{Var}_g(\mu_{g,0})$ that we penalize is upper-bounded by $\mathcal{D}_1(\theta)+\mathcal{D}_0(\theta)$, linking it to a threshold-agnostic EO notion: For any fixed threshold τ, $R_{g,1}(\tau)$ is exactly the TPR for group g (the share of Y=1 cases predicted positive at τ), and $R_{g,0}(\tau)$ is the FPR for group g (the share of Y=0 cases predicted positive at τ). The quantities $\mathcal{D}_1$ and $\mathcal{D}_0$, therefore aggregate, across all thresholds, the cross-group variability in TPR and FPR, with weights $w(\tau) = \sigma'(\tau)$ that peak near the decision boundary. Proposition 2 connects our smooth penalty to these decision outcomes by showing that big variation in our penalty cannot occur unless there is big, average disparity in TPR/FPR across decision thresholds. In other words, if the penalty is large, then the weighted EO disparity must be at least that large, so a high penalty confirms an EO problem exists, on average, near the decision boundary. In economic terms, this shows that the penalty is measuring disparities right where a lender's decisions are actually made, among applicants on the margin whose outcomes flip with small score changes. In practice, this is the region where reducing disparities has the largest impact on who gets credit and on the lender's exposure to disparate-impact concerns. In contrast, disparities far from this frontier (where everyone would be approved or everyone would be denied) have little economic impact and also small weights in the penalty bound. As Proposition 2 shows, this implicitly connects the penalty term to the practically relevant region without having to commit to a single hard threshold during training. That is, the proposition allows us to read a high penalty as confirming material cross-group differences in error rates where approvals and denials are actually determined. While, conversely, a small penalty does not guarantee small EO disparities at specific thresholds, we verify empirically that minimizing the penalty term indeed decreases EO gaps at the decision threshold below.

Third, under our sparse, additive, and smooth architecture (as characterized by the parameters below), this penalty generalizes beyond the training sample:

Proposition 3. *With a probability of at least* $1-\delta$,

$$\sup_{\theta} \left|\mathcal{P}_{\text{mean}}(\theta) - \widehat{\mathcal{P}}_{\text{mean}}(\theta)\right| = O\left(\frac{(\Lambda_1 B_1 + \Lambda_2 B_2)\sqrt{\log M} + (\Lambda_1 L_1 + \Lambda_2 L_2) + \sqrt{\log(G/\delta)}}{\sqrt{p_{\min}}\,\sqrt{n}}\right),$$

where $\Lambda_1 = \sum_i |\alpha_i|$, $\Lambda_2 = \sum_{(i,j)} |\alpha_{ij}|$ *(sparsity of main and interaction gates),* $B_1 = \sup_x |f_i(x)|$, $B_2 = \sup_x |f_{ij}(x)|$ *(amplitude bounds),* $L_1 = \mathrm{Lip}(f_i)$, $L_2 = \mathrm{Lip}(f_{ij})$ *(smoothness bounds),* M *is the number of active components,* $G = |\mathcal{G}|$ *is the number of groups, and* $p_{\min}$ *is the minimal group-label cell mass.*[7]

[7] For a precise definition of these parameters, see its proof in Appendix E.

This result is important. Its main takeaway is that the gap between the population penalty and its empirical estimate decays at the standard $O(1/\sqrt{n})$ rate, up to the minimal group-label cell mass, showing that the fairness improvement learned in training reliably carries over to new data rather than being an in-sample artifact.

In other words, Proposition 3 provides a uniform law of large numbers for our fairness metric. It shows that, with probability at least $1 - \delta$, the fairness measure computed on the sample, $\widehat{\mathcal{P}}_{\text{mean}}(\theta)$, is close to the true population value $\mathcal{P}_{\text{mean}}(\theta)$ for *all* models θ in our class. That is, the fairness improvement achieved during training is not an in-sample artifact but persists when the model is applied to new data.

Economically, this result matters because it converts a statistical property into a *reliability guarantee*. A regulator or risk committee can interpret the bound as: "with high probability, the true degree of fairness of the deployed model differs from what we measured in training only by a small, quantifiable amount." The convergence rate is the classical $1/\sqrt{n}$, adjusted for the smallest group-label cell share $p_{\min}$.[8] This scaling highlights the data requirement for dependable fairness: if a protected group has few observations in some outcome category (for instance, few minority denials), the fairness estimate will be noisier, and more data are needed for the same reliability. Thus, the term $1/\sqrt{p_{\min}}$ makes explicit how sample composition constrains credible fairness claims.

Finally, and importantly, the remaining terms in the bound summarize model complexity (and show how our model class (interpretability) helps): sparsity parameters (Λ_1, Λ_2), which count how many main and interaction effects are active; amplitude bounds (B_1, B_2), which limit how large the learned shape functions can become; and smoothness parameters (L_1, L_2), which restrict how sharply these shapes can bend. Simpler, smoother, and sparser architectures yield tighter fairness reliability, while highly complex models require larger samples to achieve the same level of assurance. Together, these quantities define a transparent *data-sufficiency condition*: given the sample size, group composition, and model complexity, Proposition 3 tells us how confidently we can expect the fairness behavior observed in training to generalize to out-of-sample deployment.

Economically, these results imply that our fairness weight λ_{EO} governs a transparent trade-off: larger λ_{EO} pushes the model toward smaller differences in group-wise error rates exactly where credit decisions are made, and the $1/\sqrt{n}$ convergence guarantee ensures that these fairness gains are not an artifact of the training sample. For regulators or risk committees, λ_{EO} therefore functions as a control parameter that trades off decision-relevant error costs against explicitly quantified fairness risk, in line with the constrained profit-maximization model in Appendix C.

[8] That is, the minimal group–label cell mass as $p_{\min} := \min_{g \in \mathcal{G},\, y \in \{0,1\}} \Pr(A = g,\ Y = y)$.

2.3 Training algorithm

The FAN model represents each f_i and f_{ij} as a small feed-forward neural network. Throughout the training process, each subnetwork is also multiplied with an additional learnable weight w_i or v_{ij}, which controls the scale of the corresponding function f_i or f_{ij}, i.e. the full definition of the score is:

$$s_\theta(X_l) \;:=\; \beta_0 \;+\; \sum_{i=1}^{p} w_i\, f_i(x_{i,l}) \;+\; \sum_{(i,j)\in I} v_{ij}\, f_{ij}(x_{i,l}, x_{j,l}).$$

While our main objective is enforcing fairness, an interesting property of this approach is that it can produce smooth, differentiable, sparse, and possibly monotone solutions.[9] To achieve this, the loss function that is reduced during training has the following structure:

$$\ell(\theta) \;=\; \underbrace{\ell^{\text{BCE}}(\theta)}_{\text{Binary cross-entropy}} \;+\; \underbrace{\lambda_{\text{EO}}\,\mathcal{P}_{\text{EO}}(\theta)}_{\substack{\text{Equalized odds penalty}\\ \text{(Fairness penalty)}}} \;+\; \underbrace{\lambda_{\text{main}} \sum_{i=1}^{p} |w_i| \;+\; \lambda_{\text{inter}} \sum_{(i,j)\in\mathcal{I}} |v_{ij}|}_{\text{LASSO-like term selection penalty}}$$
$$+\; \underbrace{\lambda_{\text{mono}}\,\mathcal{P}_{\text{mono}}(\theta)}_{\text{Penalization of mismatched signs of gradients}} \;+\; \underbrace{\lambda_{\text{clar}}\,\mathcal{P}_{\text{clar}}(\theta)}_{\text{Penalization of main-interaction covariance}}.$$

To estimate the optimal parameters θ, the training procedure implements three main stages: main (univariate) effect learning, interaction (bivariate) effect learning on the residuals of the first stage, and joint fine-tuning. We provide additional information on all components of the loss function, all training stages, and other details related to the estimation technique in Appendix D.

3 Data and Results

We now connect the modeling framework developed in Section 2 to a concrete underwriting decision: whether a mortgage application is denied or originated. The central empirical question is whether fairness-aware estimation can deliver a *less discriminatory alternative* to standard credit models— one that assigns less disparate approval and denial predictions while preserving predictive performance and remaining transparent enough to support regulator-aligned documentation and adverse-action explanations.

[9] Monotone behavior can be enforced for input variables if prior knowledge regarding their effects is available.

To answer this question, we apply FAN to U.S. mortgage lending data and evaluate its performance relative to conventional benchmarks. We focus on three outcomes: (i) group disparities in false-positive and false-negative rates, (ii) predictive accuracy, and (iii) the transparency of the resulting decision rules.

3.1 Data

Our empirical analysis uses the 2018 release of the Home Mortgage Disclosure Act (HMDA) data set, which contains detailed information on mortgage loan applications submitted to financial institutions across the United States. The raw data include more than 15 million applications, with variables describing loan characteristics, property attributes, applicant demographics, and underwriting outcomes. We use public HMDA data because it matches the institutional margin studied in this paper: origination and denial decisions at the application stage. It provides a harmonized, publicly observable, and lender-comparable record of applicant, loan, property, and neighborhood characteristics together with the corresponding underwriting outcome. This makes it especially useful for evaluating whether interpretable machine-learning models can deliver less disparate decision rules within a common information set. By design, the exercise abstracts from private underwriting variables such as credit scores and realized loan performance, which are needed for ex post default analysis but not for the public-information-set underwriting question we study here.

To obtain a clean and comparable sample from the 2018 HMDA dataset, we apply standard data filters (see, e.g., Fuster et al., 2024; Liu et al., 2019). Specifically, we restrict the data to loan applications that are (i) `lien_status` = 1 (secured by a first lien), (ii) `total_units` = 1 (single-family properties), and (iii) `occupancy_type` = 1 (principal residences). These filters ensure that the sample is more homogeneous and corresponds to the typical residential mortgage market segment, excluding second liens, multifamily loans, and investment or vacation properties.

Following Das et al. (2023), we construct a random representative sub-sample of 500,000 loans from these data. Our prediction target is whether a loan application l was *rejected* or *originated* (approved and accepted by the applicant). We construct a binary outcome variable:

$$y_l = \begin{cases} 1 & \text{if the application was denied (action_taken = 3)} \\ 0 & \text{if the application was originated (action_taken = 1).} \end{cases}$$

Other action codes (e.g. withdrawn, incomplete, purchased loans) were excluded to ensure a clean distinction between successful and rejected applications. After filtering, the

positive class (rejected) constitutes approximately 23% of the sample, while the negative class (originated) accounts for 77%.

We restrict attention to features that are plausibly available to lenders at the time of decision-making, avoiding variables that would directly encode the outcome or create leakage. For example, variables that directly reveal underwriting outcomes (`action_taken`, `denial_reason`, etc.) were excluded to avoid target leakage. Summary statistics for the selected HMDA variables are reported in Appendix A.

To evaluate fairness across protected groups, we construct a binary indicator for minority status equal to 1 if the borrower is of Black/Hispanic/Native-American/Pacific-Islander origin and 0 otherwise. This attribute is used only for loss evaluation and optimization, but not as an input feature to the predictive models.

Finally, the data are randomly split into three disjoint sets: 70% training data ($\mathcal{T}_1$) to fit models, 15% validation data ($\mathcal{T}_2$) for hyperparameter tuning, and 15% for out-of-sample testing ($\mathcal{T}_3$). We train one model for each combination within a standard grid of hyperparameters and choose the model with the highest ROC-AUC on the validation data.

3.1.1 Preprocessing

Missing values in continuous variables are imputed using the median of the training data, while missing values in categorical variables are imputed based on the most frequent value in the training data. Additionally, `loan_amount`, `income`, and `property_value` are log-transformed to mitigate skewness. We also winsorize loan-to-value ratios at the 99.9% training quantile to avoid unrealistic outliers. Before estimating the models, categorical features are transformed into dummy variables via one-hot encoding.[10] Continuous features are normalized to have means of zero and unit variance to ensure consistent scaling.

3.2 Performance evaluation

We start by training a standard logit classification model as a baseline and compare it to FAN models with varying penalty weights λ_{EO}. We evaluate classifier performance based on common metrics like accuracy and ROC-AUC. To compare algorithmic fairness, we

[10]Throughout the paper, we denote each dummy variable by the HMDA variable name and the corresponding HMDA code encoded by the dummy. For example, the HMDA field *preapproval* is represented by the two dummies `preapproval_1` (preapproval requested) and `preapproval_2` (preapproval not requested). See Appendix B for detailed variable definitions.

calculate true positive and false positive rates within groups as well as their differences across groups, in line with the equalized odds criterion.

As the results reported in Table 1 show, the machine learning model outperforms the logistic regression for all levels of penalization, improving on both performance measures. At the same time, the machine learning approach allows us to reduce fairness disparities using increasing levels of penalization without impeding accuracy or ROC-AUC and while consistently beating the logit benchmark. Figure 1, which compares within-group ROC curves for the two FAN models with $\lambda_{EO} = 0$ and $\lambda_{EO} = 3$, shows that the penalized model achieves its fairness improvements without sacrificing predictive performance in terms of AUC. In other words, the overall pattern of errors that would be costly under standard profit motives does not materially worsen, even as the model becomes much more equal across groups.

While we confirm the general observation of Das et al. (2023) that group disparities, overall, are moderate to small in HMDA rejection rates, note that this approach can be applied to other classification tasks in consumer lending, where significant disparities between groups and/or machine learning models have been documented, such as loan interest rates (Bartlett et al., 2022b) or default rates (Fuster et al., 2022).

	Accuracy	ROC–AUC	$\lvert \text{TPR}_1 - \text{TPR}_0 \rvert$	$\lvert \text{FPR}_1 - \text{FPR}_0 \rvert$
Logit (baseline)	0.829	0.802	0.065	0.014
FAN ($\lambda_{\text{EO}} = 0$)	0.838	0.823	0.059	0.006
FAN ($\lambda_{\text{EO}} = 1$)	0.838	0.823	0.051	0.005
FAN ($\lambda_{\text{EO}} = 2$)	0.840	0.824	0.050	0.004
FAN ($\lambda_{\text{EO}} = 3$)	0.839	0.823	0.043	0.003

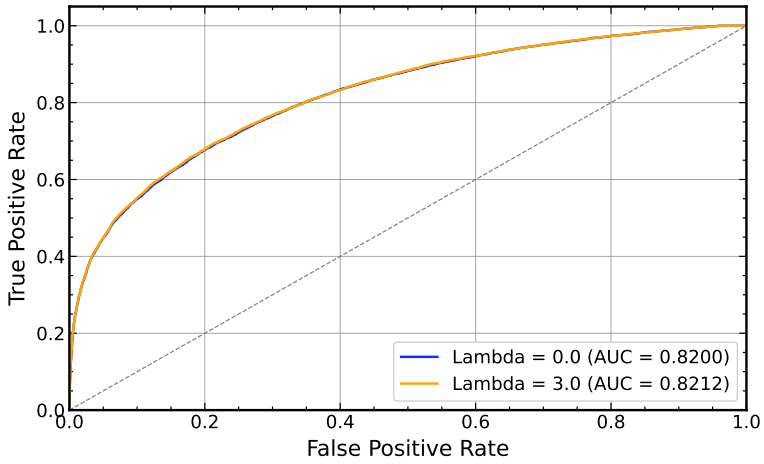
Table 1: Performance and fairness metrics for baseline Logit and FAN models on the test data. The fairness metrics are reported as absolute differences in group-specific rates between minority ($a = 1$) and non-minority ($a = 0$).

3.3 Predictor importance

In the following, we make use of the interpretable nature of our model in order to identify the most important effects. We define the *mean absolute score* $S(i)$ as the absolute values of $f_i(x_{i,l})$ averaged over the training sample:

$$S(i) := \frac{1}{|\mathcal{T}_1|} \sum_{l \in \mathcal{T}_1} |f_i(x_{i,l})|. \tag{2}$$

A: Non-minority group



B: Minority group

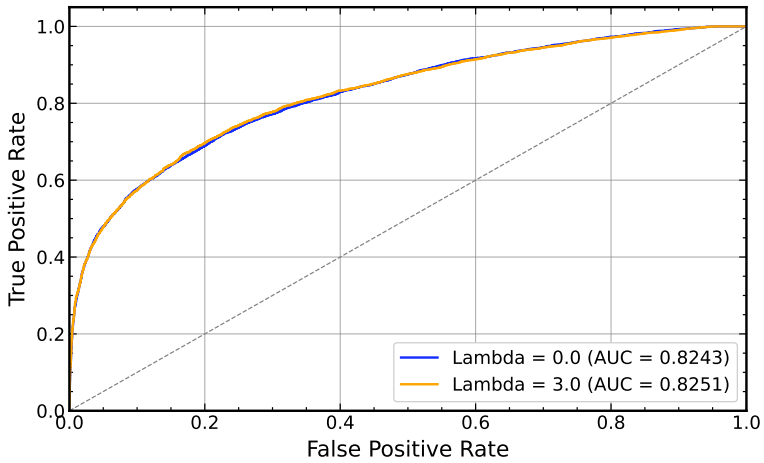


Figure 1: Group ROC curves. ROC curves are calculated separately for non-minority (Panel A) and minority applicants (Panel B) on the test data. Results are reported for the unconstrained baseline model ($\lambda_{EO} = 0$) and the most heavily penalized model ($\lambda_{EO} = 3$), for comparison.

$S(i)$ captures the average effect of variable i, taking the distribution over the sample into account, and is therefore similar to the absolute value of a (standardized) coefficient in a linear regression.[11] Hence, the mean absolute scores capture the average "size" of the effect of an individual variable on the prediction of the model. The mean absolute scores of the interaction terms are defined analogously:

$$S(i,j) := \frac{1}{|\mathcal{T}_1|} \sum_{l \in \mathcal{T}_1} |f_{ij}(x_{i,l}, x_{j,l})|. \tag{3}$$

Figure 2 displays the resulting rankings for the 15 most important prediction terms (univariate functions and interactions) in the cases $\lambda_{EO} = 0$ and $\lambda_{EO} = 3$. Importance scores $S(i)$ of univariate effects are identified using the name of the corresponding variable x_i, while interactions are denoted by "x_i x x_j".

Across both specifications, the top predictors are mostly dominated by *univariate* effects. For λ_{EO}=0, the two largest contributors are `loan_purpose_1` and `dti_num`, followed (at some distance) by `log1p(loan_amount)` and `loan_to_value_ratio`.[12] The leading interactions then enter, namely `log1p(loan_amount)` × `log1p(property_value)` and `dti_num` × `loan_type_1`, along with contextual and demographic terms such as `tract_minority_population_percent`, `ffiec_msa_md_median_family_income`, and age bins. Additional interactions appear lower in the list (e.g., `preapproval_1` × `construction_method_2`, `loan_term` × `loan_purpose_1`).

With fairness regularization (λ_{EO}=3), the overall ordering is remarkably stable: `loan_purpose_1` and `dti_num` remain the top two, and `log1p(loan_amount)` and `loan_to_value_ratio` stay in the top four. The same two interactions, `log1p(loan_amount)` × `log1p(property_value)` and `dti_num` × `loan_type_1`, continue to be the most important pairwise terms, while others shift: `loan_term` × `loan_purpose_1` persists, and `preapproval_2` × `mfd_home_sec_prop_type_2` and `loan_term` × `ddc_sf_site_built` newly enter the top fifteen. Notably, some geographic proxies lose relevance under the penalty: `tract_minority_population_percent` (present at λ=0) drops out, while core underwriting determinants (loan purpose, DTI, LTV, and loan amount) retain their prominence. Finally, `log1p(income)` enters the top 15 under λ=3, reflecting a modest shift toward direct repayment capacity measures.

[11] In a linear model with standardized inputs, the contribution of variable i to the score is $\beta_i x_{i,l}$. So our "importance" summary is the average absolute contribution: $\frac{1}{n}\sum_l |\beta_i x_{i,l}|$. Standardization puts x_i on a comparable scale across variables, so $\frac{1}{n}\sum_l |\beta_i x_{i,l}|$ is (up to the factor $\frac{1}{n}\sum_l |x_{i,l}|$, which tends to be similar across standardized variables) proportional to $|\beta_i|$.

[12] `log1p` denotes our log transformation $\text{log1p}(x) = \ln(1+x)$.

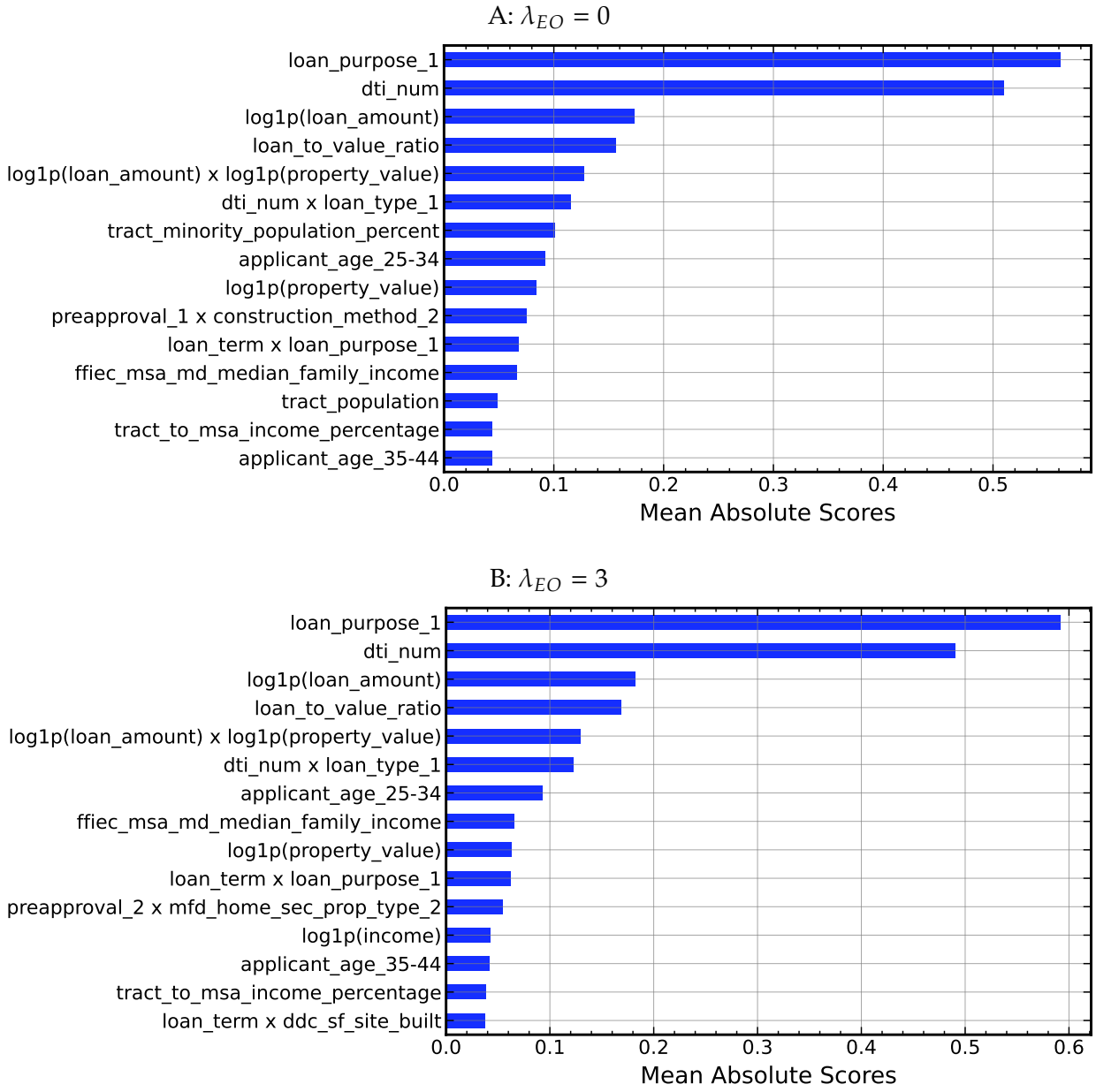


Figure 2: Mean Absolute Scores $S(i)$ and $S(i, j)$. This figure compares the most important predictors for the FAN models with $\lambda_{EO} = 0$ (Panel A) and $\lambda_{EO} = 3$ (Panel B).

3.4 Shape functions

Due to the separable structure of the FAN model, we can extract the output of each subnet corresponding to an individual input characteristic or interaction term. This allows us to interpret the shape functions f_i and f_{ij} as the global influence of each respective term in equation 1, enabling model explanation in line with regulatory requirements. In contrast to traditional machine learning models, these functions capture the influence of each model term *without* approximation error from post-hoc explainability methods (see, e.g., Rudin, 2019). In the following, we focus on the most important univariate effects first and discuss interactions in Section 3.5.

Figure 3 presents the shape functions f_i corresponding to the four most important univariate predictors. To show the impact of the adjusted loss function, we compare the unconstrained model ($\lambda_{EO} = 0$) with a penalized model ($\lambda_{EO} = 3$). We find a number of non-linear functions that determine rejection/origination probabilities.

The effect of `loan_purpose_1`, a dummy variable encoding whether the loan is for home purchase or not[13], is presented in Panel A. For home purchase loans, the corresponding shape function decreases the model output, while loans for other purposes are penalized, i.e., experience an increase in rejection probability. This lower rejection risk for purchase loans could reflect the fact that refinances and cash-out loans tend to extract equity and raise borrower leverage, making them riskier and less favorable to lenders. Panel B illustrates the impact of the debt-to-income ratio (dti)[14], which results in monotonically increasing additions to the prediction score, consistent with deeply indebted borrowers being less attractive to lenders. High dti values raising rejection risk is intuitive as they signal limited repayment capacity and higher probability of financial distress. The sharper increase at high dti levels reflects nonlinear risk: once borrowers cross common underwriting thresholds, even small additional debt burdens sharply worsen default expectations. In contrast, Panel C, which shows the univariate impact of `loan_amount`, indicates that the optimal shape functions need not be monotonic. The function increases from small negative values around the mean to large positive contributions in the tails. The non-monotonic effect could reflect two offsetting forces: moderate loan sizes are routine and relatively safe, while very small loans may indicate distressed properties or limited borrower resources, and very large loans amplify loss severity in default. The model captures such a U-shaped risk profile, penalizing both extremes more heavily than mid-range loans. Panel D shows the effect associated with an application's `loan-to-value ratio`

[13]Other purposes include home improvement, refinancing, and cash-out refinancing.

[14]The variable `dti_num` is the categorical HMDA variable `debt_to_income_ratio` transformed to a numeric variable based on bin centers.

(LTV). The estimated shape function is slightly negative for below-mean LTVs, crosses to a small positive value at the mean, and then remains roughly flat until about one standard deviation above the mean. Beyond that point the function increases sharply and approximately linearly, reaching values around 4 at the highest LTVs. Economically, this pattern is intuitive: at low leverage, substantial borrower equity reduces expected loss given default and decreases collateral risk, leading to lower rejection scores. Around typical leverage levels, marginal changes in LTV carry limited incremental information once other covariates are held fixed, so the effect is close to zero and nearly flat. At high LTVs, low borrower equity raises expected losses and valuation uncertainty, may trigger stricter underwriting requirements (e.g., mortgage insurance or program cutoffs), and increases the perceived risk of borrower default, resulting in a steep increase in the rejection score.

Additionally, the comparison between the two models illustrates how the predictive process needs to be adjusted to account for the fairness objective. In particular, we observe higher increases in rejection probabilities for non-home purchase loans (`loan_purpose` $\neq$ 1) when group disparities are penalized, suggesting that purpose might be correlated with protected-group status. Additionally, the fairness adjustment dampens the overall impact of high debt-to-income ratios on the classification output, with slightly lower functions values and a reduced slope for above-average dti outcomes. Furthermore, high (low) loan amounts result in lower (higher) increases relative to the unconstrained model, suggesting that disproportionate approvals (denials) at the lower (higher) end may have contributed to group disparities. For the LTV shape function, the EO penalty leaves the functional form essentially unchanged across specifications. The curves for $\lambda = 0$ and $\lambda = 3$ nearly coincide over the entire LTV range, with only a slight divergence at low LTVs: the EO-constrained model produces marginally more negative values (i.e., slightly lower rejection scores) for safer, low-LTV applications. This suggests that the fairness objective does not materially alter the risk-gradient captured by LTV, while inducing a small additional preference toward strongly collateralized loans.[15]

3.5 Interactions

In Figure 4, we compare two of the most important interactions that appear in both the unconstrained model and the penalized model. Panels A and B show interaction heatmaps with the `loan_amount` on the x-axis and the value of the property proposed to secure the covered loan (`property_value`) on the y-axis. The color-coded third axis represents the

[15] In other words, $\lambda = 3$ *doesn't rewrite underwriting logic*, but it tilts certain margins (purpose, high DTI, loan-size extremes) in ways consistent with reducing EO disparities, while leaving core risk gradients like LTV largely intact.

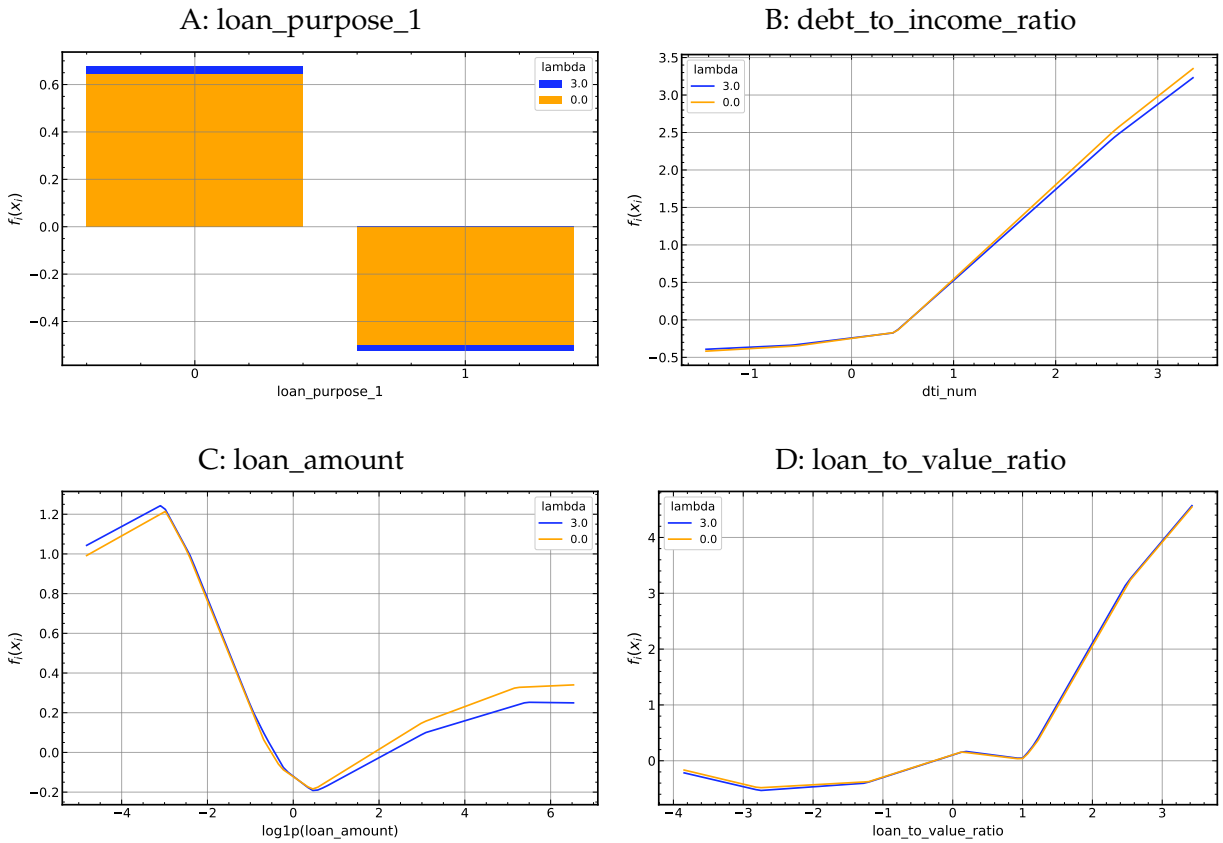


Figure 3: Univariate shape functions $f_i(x_i)$. For numerical features, the x-axis plots the input variable x_i standardized to a mean of 0 and a standard deviation of 1. The y-axis represents the additive contribution to the model score $s(X)$ through the shape function $f_i(x_i)$.

corresponding raw function output f_{ij}. The interaction surface is highly asymmetric, but the pattern is economically clear. In the upper-left region (small loans and relatively high property values), the interaction is close to zero, so this pair adds little to the rejection score once main effects are held fixed. Crossing the near-diagonal boundary toward the lower-right region (larger loans relative to collateral), the interaction becomes strongly positive and rises quickly, adding a large rejection-score penalty. It also forms a sharp ridge in the extreme corner where `log1p(loan_amount)` is high while `log1p(property_value)` ranges from about −6 to +3 standard deviations. For $\lambda = 0$, the surface peaks around +12; for $\lambda = 3$, the shape is very similar but the peak increases to about +15 and is more tightly concentrated. This pattern is consistent with an additional LTV-like effect concentrated on the most risky loans: Small loans against relatively high-value properties (upper-left triangle) present minimal incremental risk from the interaction channel, leaving the predicted effect close to zero. By contrast, when the requested loan amount is large and the securing property is relatively less valuable, implied leverage is high and the interaction term rises sharply to a large rejection score that dominates other shape function outputs in comparison. Economically, this is consistent with thinner collateral increasing expected losses and required capital, so underwriting decisions become more conservative and the rejection score rises.

Panels C and D show the heatmap plots for the interaction of `dti_num` on the x-axis and the dummy for `loan_type` = 1 (conventional loan not insured or guaranteed by FHA, VA, RHS, or FSA)[16] on the y-axis. The interaction term has its largest effects for moderately high debt-to-income ratios (dti), slightly more than one standard deviation above the mean, with considerable heterogeneity across loan types. For conventional loans, higher dti raises the rejection score while lower dti reduces it. For insured/guaranteed loans, the pattern is reversed. This interaction illustrates how lenders evaluate repayment risk differently across loan types: relatively high dti sharply raises rejection risk for conventional loans, while government-backed programs mitigate that risk by providing insurance against default. Conversely, low-dti applicants may face slightly higher rejection rates in insured products, consistent with program design that targets higher-risk borrowers. In addition, we document a slight shift between the unconstrained and the penalized model, resulting in more pronounced decreases in rejection probabilities for insured/guaranteed loans when fairness is taken into account.

[16]Other loan types are Federal Housing Administration insured (FHA), Veterans Affairs guaranteed (VA), USDA Rural Housing Service or Farm Service Agency guaranteed (RHS or FSA).

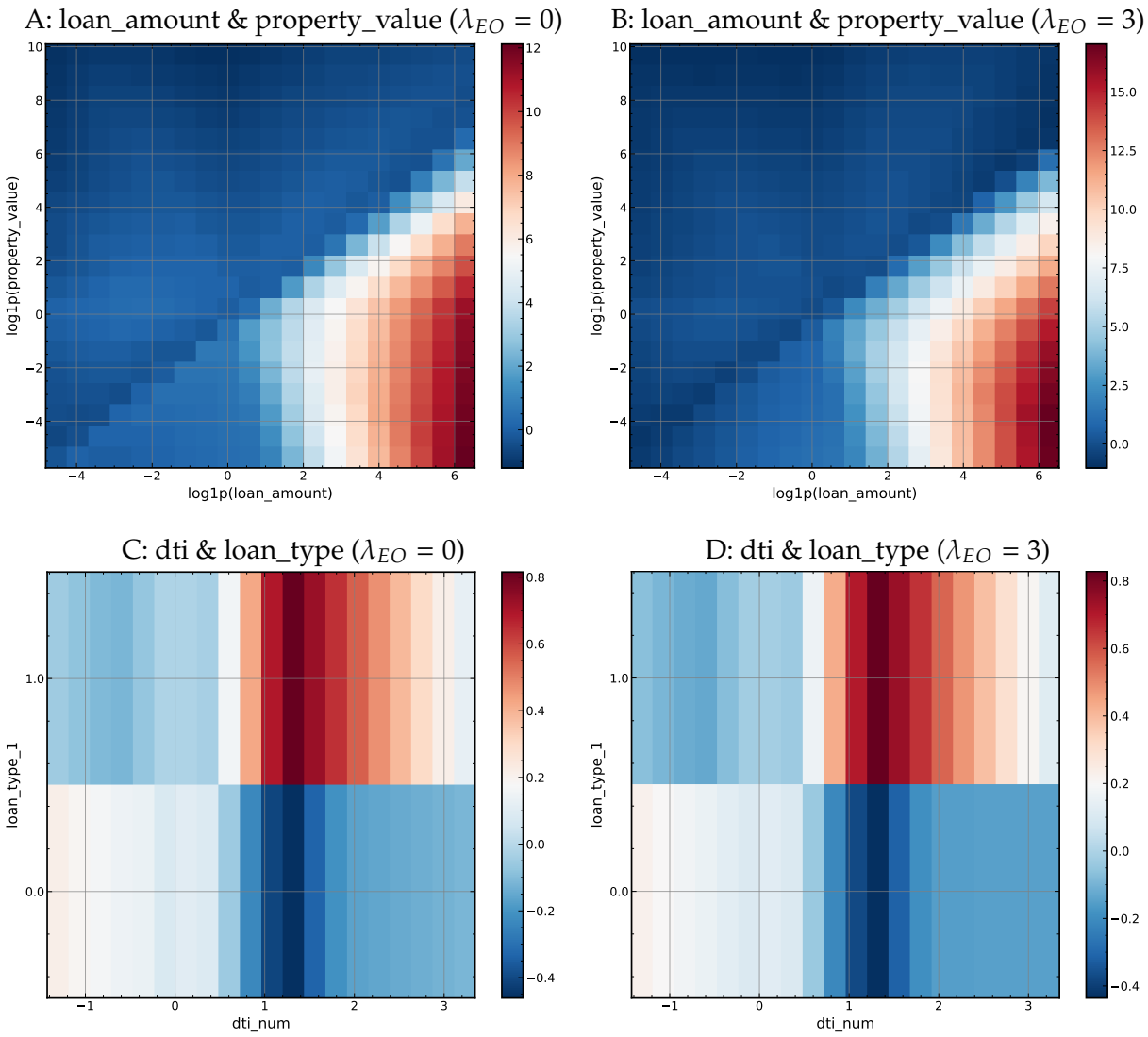


Figure 4: Interaction effects $f_{ij}(x_i, x_j)$. The scaled values of x_i and x_j are shown on the x- and the y-axis, respectively. The color-coded axis represents the additive contribution to the model score associated with combinations of x_i and x_j values.

4 Behavior at underwriting thresholds

So far, our empirical results have shown sizable minority-specific disparities in error rates (relative to observed denial/origination labels) and suggested that fairness regularization can shrink those gaps while retaining overall model performance. However, do these results reflect real differences in how similar applicants are treated rather than model misspecification or compositional noise? To answer this question, we leverage a discontinuity at the LTV = 80% underwriting threshold, a well-established boundary for requiring private mortgage insurance (PMI) and other high-leverage underwriting rules, to implement a sharp regression discontinuity design. The relevance of this LTV threshold has been observed in previous literature (Adelino et al., 2025; Green and Wachter, 2005). Our objective is to show how our models act as a counterfactual by studying whether approval decisions exhibit discontinuities at the LTV = 80% boundary and whether these discontinuities differ by minority status.[17]

For loan application l, let the running variable be $r_l := \text{LTV}_l - 80$ (in percentage points), and $D_l := \mathbb{1}\{r_l \geq 0\}$ indicate being (weakly) above the threshold. Let $A_l \in \{0, 1\}$ be the minority indicator defined in Section 3.1. On the test data, we estimate a local linear RD with triangular kernel weights $w_l(h) := \max\{0, 1 - |r_l|/h\}$ in the window $|r_l| \leq h$:

$$Y_l \;=\; \alpha + \beta\, r_l + \gamma\, D_l + \theta\, A_l + \zeta\,(A_l \times D_l) \;+\; \eta\,(D_l \times r_l) \;+\; \varepsilon_l, \tag{4}$$

estimated by weighted least squares with heteroskedasticity-consistent standard errors. The coefficient γ is the discontinuity (jump) in approval probability for the baseline group with $A = 0$ (non-minority applicants) at r=0. The minority differential at the cutoff is ζ, so the minority jump equals γ+ζ. We report results for bandwidths $h \in \{3, 5, 10\}$ percentage points (p.p.).

Table 2 reports the estimates with a bandwidth of h=10 p.p.[18] For model-generated approvals, we find a sizable and precisely estimated negative jump in approval probability at LTV≈80%: $\hat{\gamma}$ ranges from about −6.8 to −4.8 p.p. across λ, all statistically significant, consistent with the underwriting rule reducing approvals just above the PMI boundary. The minority differential at the boundary is also large and negative, with $\hat{\zeta}$ between

[17] In the following, $Y_l = 1$ represents approval.

[18] To be more clear we estimate the same RD specification in Eq. (4) for (i) observed approval decisions in HMDA and (ii) model-implied approvals on the test set for each fairness weight λ_{EO}. The observed RD confirms that the institutional cutoff generates a discontinuity in realized decisions, while the model-based RD treats the trained classifier as a counterfactual decision rule holding the applicant pool fixed–so differences in the minority discontinuity isolate how the decision rule changes treatment at the underwriting margin.

about −7.7 and −6.9 p.p., implying that conditional on being close to the boundary at LTV= 80%, minority applicants experience a materially lower approval jump than otherwise similar non-minority applicants. Figure 5 displays local linear fits on each side of the cutoff, illustrating the discontinuity. Observed approvals in HMDA show the same sign pattern but with smaller magnitudes: $\hat{\gamma} = -2.31$ p.p. and $\hat{\zeta} = -3.07$ p.p., both statistically significant. Crucially, we find that the minority $\hat{\zeta}$ declines in magnitude as the fairness weight increases: from −7.74 p.p. at λ=0 to −6.86 p.p. at λ=3.

Table 3 varies the bandwidth $h \in \{3, 5, 10\}$ to assess the local sensitivity of these findings. For model approvals, the negative jump persists and remains statistically significant across all h and λ: $\hat{\gamma}$ lies between about −7.7 and −4.8 p.p. The minority differential $\hat{\zeta}$ is consistently negative and highly significant across all bandwidths and fairness weights, with magnitudes clustered between about −7.6 and −6.1 p.p. Importantly, $\hat{\zeta}$ becomes less negative as the fairness penalty grows. At h=3, $\hat{\zeta}$ moves from −7.01 to −6.11 p.p. between λ=0 and λ=3, at h=5, from −7.61 to −6.66 p.p., and at h=10, from −7.74 to −6.86 p.p. While the sequence is not strictly monotone at every intermediate λ, the overall pattern is a clear decrease in the minority-specific discontinuity as λ increases. For observed approvals, $\hat{\gamma}$ is again negative and significant for all h, while $\hat{\zeta}$ becomes more negative and more precisely estimated as h widens (not significant at h=3, statistically significant and economically meaningful at h=5 and 10 with −1.97 and −3.07 p.p., respectively).

Overall, the RD evidence therefore supports the interpretation that fairness regularization meaningfully shrinks minority-specific discontinuities at underwriting thresholds while retaining the broad structure of rule-driven approval patterns.[19] Fairness regularization directly reduces disparities on the margin where credit is rationed. This is precisely the region that matters for access to credit and for disparate-impact exposure, since small shifts in the score around these cutoffs can flip approvals without changing overall capacity.

5 Robustness and additional analysis

In order to demonstrate the robustness of our results and illustrate the underlying economic mechanisms, we study several extensions and related research questions. In Section 5.1, we analyze adverse action reasons by identifying which variables most often push applications toward denial and comparing these channels under fairness penalization. Section 5.2 studies which predictors explain cross-group differences and how the fair-

[19] In Appendix H, we confirm that this pattern persists when applying a similar RD design at the decision boundary at DTI=43% and provide additional placebo and robustness tests.

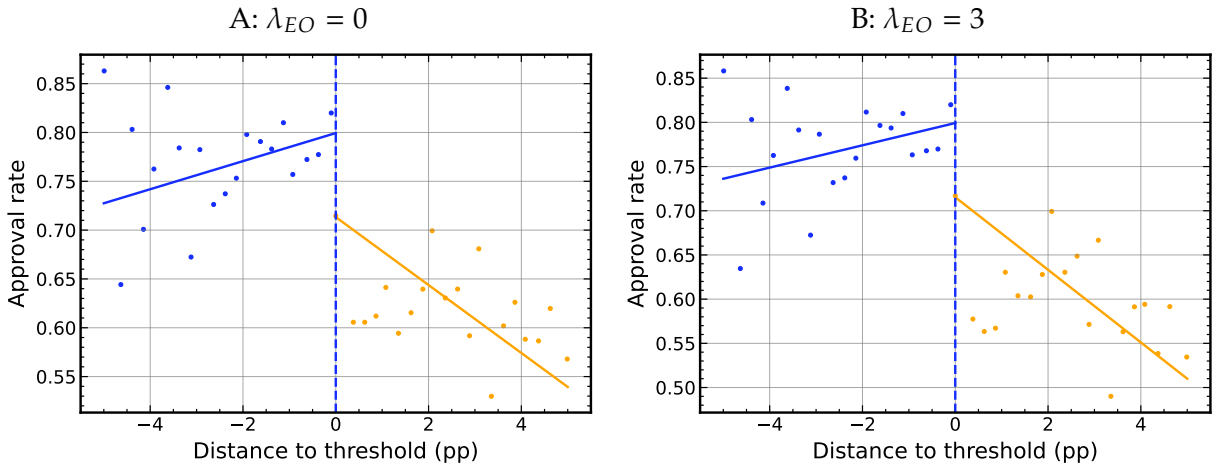


Figure 5: Local linear fit at LTV=80% for model approvals (h=5). The figure shows loan-to-value ratios and approval rates for the unconstrained FAN (Panel A) and the penalized model (Panel B) with the corresponding regression lines below and above the 80% threshold. Each dot represents mean LTV and mean approval within one of 40 bins of test set observations.

Outcome	N	h	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\zeta}$
Model approvals, λ=0	31463	10.0	0.9129***	-0.0648***	-0.0774***
			(0.0066)	(0.0074)	(0.0140)
Model approvals, λ=1	31463	10.0	0.9123***	-0.0635***	-0.0725***
			(0.0066)	(0.0074)	(0.0140)
Model approvals, λ=2	31463	10.0	0.9110***	-0.0483***	-0.0723***
			(0.0066)	(0.0073)	(0.0139)
Model approvals, λ=3	31463	10.0	0.9128***	-0.0682***	-0.0686***
			(0.0066)	(0.0074)	(0.0139)
Observed approvals	209977	10.0	0.7870***	-0.0231***	-0.0307***
			(0.0037)	(0.0040)	(0.0068)

Table 2: Regression results around LTV= 80% with bandwidth h=10 p.p. This table reports intercepts and discontinuities estimated from the RD design specified in (4). N is the size of the estimation sample taken from the test data for model approvals and from the full data for observed approvals. Significance levels are denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

h	N	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\zeta}$
		Model approvals, λ=0		
3.0	16710	0.9222***	-0.0729***	-0.0701***
		(0.0103)	(0.0109)	(0.0224)
5.0	22536	0.9230***	-0.0737***	-0.0761***
		(0.0086)	(0.0093)	(0.0191)
10.0	31463	0.9129***	-0.0648***	-0.0774***
		(0.0066)	(0.0074)	(0.0140)
		Model approvals, λ=1		
3.0	16710	0.9213***	-0.0712***	-0.0655***
		(0.0104)	(0.0109)	(0.0224)
5.0	22536	0.9228***	-0.0727***	-0.0711***
		(0.0087)	(0.0093)	(0.0191)
10.0	31463	0.9123***	-0.0635***	-0.0725***
		(0.0066)	(0.0074)	(0.0140)
		Model approvals, λ=2		
3.0	16710	0.9236***	-0.0589***	-0.0662***
		(0.0101)	(0.0107)	(0.0225)
5.0	22536	0.9218***	-0.0574***	-0.0724***
		(0.0086)	(0.0092)	(0.0191)
10.0	31463	0.9110***	-0.0483***	-0.0723***
		(0.0066)	(0.0073)	(0.0139)
		Model approvals, λ=3		
3.0	16710	0.9214***	-0.0756***	-0.0611***
		(0.0103)	(0.0109)	(0.0224)
5.0	22536	0.9223***	-0.0766***	-0.0666***
		(0.0086)	(0.0093)	(0.0191)
10.0	31463	0.9128***	-0.0682***	-0.0686***
		(0.0066)	(0.0074)	(0.0139)
		Observed approvals		
3.0	111406	0.8185***	-0.0509***	-0.0101
		(0.0057)	(0.0059)	(0.0111)
5.0	150550	0.8023***	-0.0356***	-0.0197**
		(0.0048)	(0.0051)	(0.0094)
10.0	209977	0.7870***	-0.0231***	-0.0307***
		(0.0037)	(0.0040)	(0.0068)

Table 3: Regression results around LTV=80% across bandwidths $h \in \{3, 5, 10\}$. This table reports intercepts and discontinuities estimated from the RD design specified in (4) across bandwidths. N is the size of the estimation sample taken from the test data for model approvals and from the full data for observed approvals. Significance levels are denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

ness penalty reallocates those differences. In Section 5.3, we examine disparities on the lender-level. Section 5.4 studies the impact of fairness regularization on risk parity and pricing alignment within the subset of approved loans. Further robustness exercises are conducted in Appendices G–M.

5.1 Adverse action reasons

So far, we have demonstrated the capabilities of the FAN model in terms of classification performance and fairness improvements. Next, we show how the model provides insights into adverse action reasons. Particularly, we focus on the subset of rejected applications and measure which model terms are responsible for increasing the output signal, pushing applications toward high rejection probabilities.

Let $\mathcal{P}_y \subset \mathcal{T}_1$ be the set of training-sample indices l that fall into class $Y = y \in \{0, 1\}$. To isolate signed contributions, we decompose each shape function output $z = f_i(x_{i,l,t})$ (for some feature i), or $z = f_{ij}(x_{i,l,t}, x_{j,l,t})$ (for some pair (i, j)) into its positive and negative part:

$$(z)_+ := \max\{z, 0\}, \qquad (z)_- := \max\{-z, 0\}.$$

We define *directional* mean scores on any class y as the class size-normalized averages of the positive and negative parts of the shape functions:

$$S_i^+(y) := \frac{1}{|\mathcal{P}_y|} \sum_{l \in \mathcal{P}_y} \big(f_i(x_{i,l,t})\big)_+ \tag{5}$$

$$S_{ij}^+(y) := \frac{1}{|\mathcal{P}_y|} \sum_{l \in \mathcal{P}_y} \big(f_{ij}(x_{i,l,t}, x_{j,l,t})\big)_+ \tag{6}$$

$$S_i^-(y) := -\frac{1}{|\mathcal{P}_y|} \sum_{l \in \mathcal{P}_y} \big(f_i(x_{i,l,t})\big)_- \tag{7}$$

$$S_{ij}^-(y) := -\frac{1}{|\mathcal{P}_y|} \sum_{l \in \mathcal{P}_y} \big(f_{ij}(x_{i,l,t}, x_{j,l,t})\big)_-. \tag{8}$$

By construction, S^+ summarizes how strongly (on average) a feature or interaction contributes positively within the chosen class, whereas S^- summarizes the average negative contribution. For our adverse action analysis we focus on scores pushing the model output upwards toward rejection

$$S_i^+ := S_i^+(1) \quad \text{and} \quad S_{ij}^+ := S_{ij}^+(1).$$

Note that this step, once again, relies on the separable model structure, which allows for exact decomposition of the model output (the classification score) into term-level contributions (the shape function outputs).

The resulting importance rankings are presented in Figure 6. While these represent aggregate contributions to adverse decisions (rejections), note that a similar decomposition can be applied locally to explain individual loan decisions, aligning with regulatory requirements. Across both the unconstrained ($\lambda = 0$) and fairness-regularized ($\lambda = 3$) models, the leading source of upward movements in the denial score is `dti_num`, the debt-to-income (DTI) ratio. In the unconstrained model, `dti_num` contributes the largest mean positive score (2.30), indicating that high leverage/limited cash flow remains the primary channel through which applications are driven toward rejection. Economically, higher DTI decreases borrowers' repayment capacity and raises default risk, so the model assigns a large positive contribution to the denial score as DTI increases. The interaction `dti_num` × `loan_type_1` (conventional) is also material (0.26), consistent with dependence of DTI's effect on the existence of insurance or guarantees relative to conventional loans.

Programmatic and size effects are the next most important contributors. `loan_purpose_1` (purchase) exhibits a sizable positive mass (0.64), with additional positive contributions from `loan_purpose_2` and `loan_purpose_4` (0.35 and 0.29, respectively). These patterns are consistent with underwriting frictions that vary by purpose, possibly, for example, due to differences in documentation intensity, equity extraction, and appraisal complexity, pushing marginal loan applications toward denial in certain purpose categories. Loan size, captured by `log1p(loan_amount)`, adds further positive mass (0.40), reflecting the sometimes large positive effects toward both ends of the size distribution documented in Section 3.4.

Collateral-related variables enter both through local market conditions and the previously discussed interaction term (see Section 3.5). The variable `ffiec_msa_md_median_family_income` contributes sizably (0.27), suggesting that neighborhood income might proxy for broad affordability and market depth that influence recoveries. The interaction `log1p(loan_amount)` × `log1p(property_value)` (0.21) indicates that the denial signal intensifies for cases where large requested balances coincide with lower relative property valuation, consistent with the steep interaction shape documented above. Two additional terms complete the top ranks at λ=0: a regional indicator, `state_code_ND` (0.25), which likely captures localized market structure and policy features not otherwise absorbed, and `preapproval_1` × `construction_method_2` (0.17), which links process features with build type, possibly adding denial contribution where screening or valuation may be more complex.

Under fairness regularization (λ=3), the model preserves the same economic drivers and reweights them modestly. `dti_num` remains the dominant contributor (2.22), and `dti_num` × `loan_type_1` stays among the largest interactions (0.26), again signaling varying capacity constraints for conventional vs. guaranteed loans at the margin. Purpose effects become slightly more prominent: `loan_purpose_1` rises (0.68), with `loan_purpose_2` and `loan_purpose_4` also increasing (0.38 and 0.33). Size and collateral continue to matter through `log1p(loan_amount)` (0.40), `ffiec_msa_md_median_family_income` (0.26), and the `loan_amount- property_value` interaction (0.21).

Two notable reallocations emerge. First, the interaction associated with `preapproval` that appeared at λ=0 is no longer among the top contributors at λ=3, suggesting that denial mass shifts away from process-related frictions and toward core economic levers (capacity, purpose, size, and collateral). Second, `applicant_age_8888` (missing/withheld age category) enters as a sizable positive contributor (0.69), indicating that missingness on key demographic inputs is associated with higher predicted risk, plausibly by proxying for informational frictions or atypical borrower characteristics. The regional indicator `state_code_ND` remains present (0.27), again suggesting residual geographic heterogeneity in recoveries or program take-up.

Taken together, these decompositions point to a stable decision logic across models, while showing which channels to prioritize in order to achieve approval: Overall, the most effective paths for marginal applicants to convert denials into approvals are economically intuitive and actionable: improve DTI (e.g., via income verification, debt reduction, or lower payments), manage balance size (smaller requested loan or higher down payment), and improve collateral position (thereby strengthening recoveries). Interactions play a role as well, especially the exposure-collateral term tying loan size to property value, while fairness regularization reallocates some adverse-action mass away from process-driven interactions toward these core channels, without overturning the underlying economics.

5.2 Decomposition of the EO disparity

In a related exercise, we investigate which features generate the EO disparity in our models and how λ_{EO} changes that composition. The goal is to attribute cross-group differences in the model's decision boundary to individual features of HMDA loan applications in a way that is exact for our additive architecture and comparable across fairness weights. Recall that $s_\theta(x)$ denotes the model score and $\hat{p}_\theta(x) = \sigma\big(s_\theta(x)\big)$ the predicted denial probability. In order to explain the observed disparities in our models' predictions, we are interested

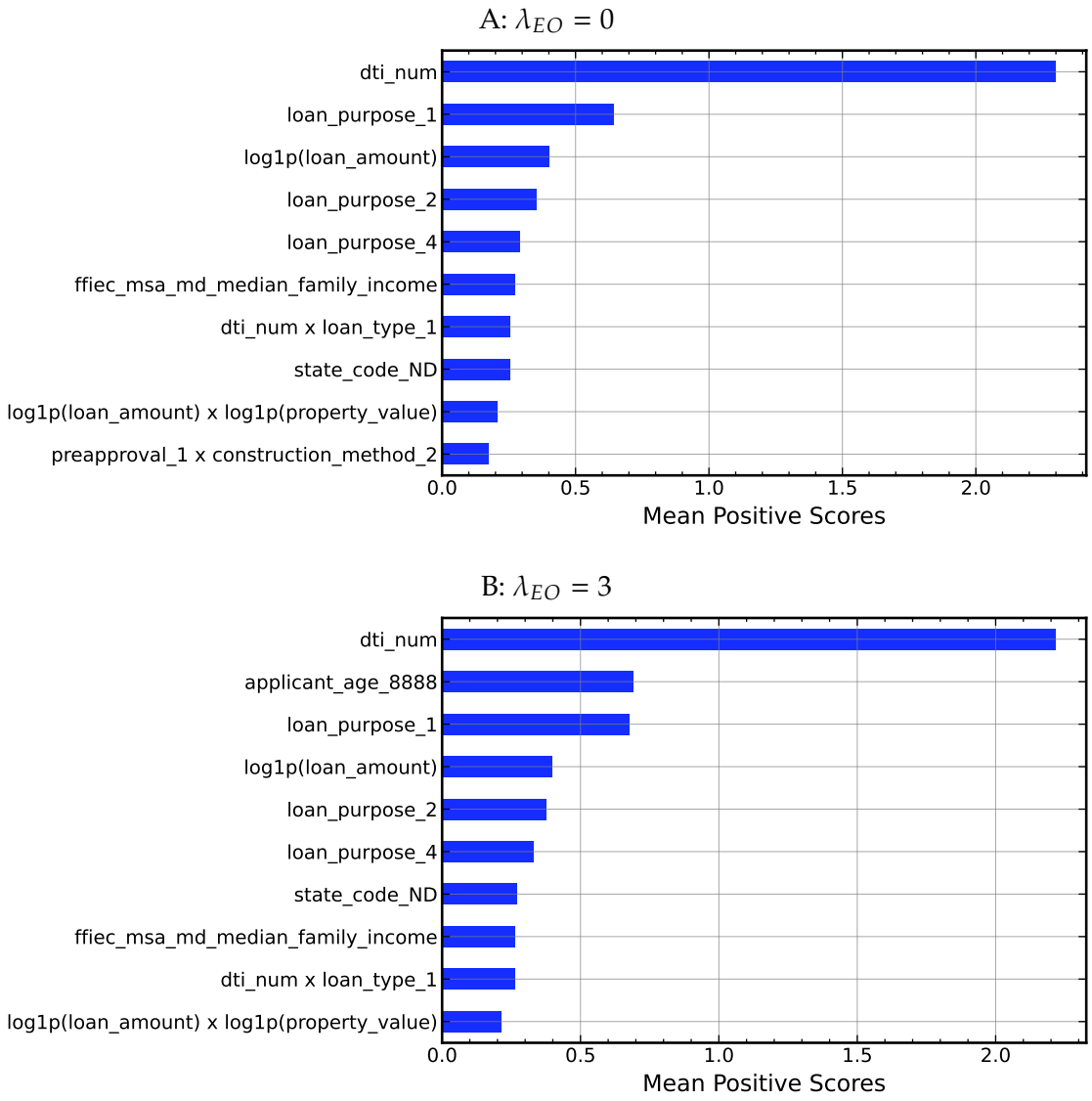


Figure 6: Mean positive scores S^+ (top ten). This figure compares the most important contributors to rejections for the FAN models with $\lambda_{EO} = 0$ (Panel A) and $\lambda_{EO} = 3$ (Panel B).

in the quantity

$$\mathbb{E}\big[\sigma(s_\theta(X)) \mid A{=}1,\, Y{=}y\big] - \mathbb{E}\big[\sigma(s_\theta(X)) \mid A{=}0,\, Y{=}y\big]\,,$$

which captures group-conditional differences in output probabilities. To now apply an additive decomposition provided by our separable model and investigate drivers of disparities at the decision boundary where the fairness penalty acts (see Section 2.2), we take advantage of the properties of the sigmoid function: The logistic link $\sigma(\tau)$ is strictly increasing and close to linear in a neighborhood of the decision threshold $\tau = 0$, so, within this region, changes in class-conditional mean scores $\mathbb{E}[s_\theta(X) \mid A, Y]$ are directionally and approximately proportional to changes in class-conditional mean probabilities $\mathbb{E}[\sigma(s_\theta(X)) \mid A, Y]$.[20] Locally, we can, therefore, write

$$\begin{aligned}&\mathbb{E}\big[\sigma(s_\theta(X)) \mid A{=}1,\, Y{=}y\big] - \mathbb{E}\big[\sigma(s_\theta(X)) \mid A{=}0,\, Y{=}y\big]\\ &\quad\propto\ \mathbb{E}\big[s_\theta(X) \mid A{=}1,\, Y{=}y\big] - \mathbb{E}\big[s_\theta(X) \mid A{=}0,\, Y{=}y\big]\,.\end{aligned}$$

This allows us to reason about the output probabilities based on the separable structure of our model on the score level. For any main or interaction term $t \in \{i\} \cup \{(i,j)\}$, we can now utilize the corresponding shape function f_t and define

$$\bar{f}_t^{\,a,y} \;:=\; \mathbb{E}\big[f_t(X)\big|A{=}a,\, Y{=}y\big]\,, \qquad \Delta_t^{(y)} \;:=\; \bar{f}_t^{\,1,y} \;-\; \bar{f}_t^{\,0,y}.$$

These represent a decomposition of the *class-conditional mean score gap* between groups into per-term contributions, as summing over all additive terms yields the group difference in the mean score within label y:

$$\mathbb{E}\big[s_\theta(X) \mid A{=}1,\, Y{=}y\big] - \mathbb{E}\big[s_\theta(X) \mid A{=}0,\, Y{=}y\big] \;=\; \sum_t \Delta_t^{(y)}.$$

The vector $\{\Delta_t^{(y)}\}_t$ is therefore an exact attribution of the class-conditional score gap into economically interpretable term-wise components.

Figure 7 presents the results for test data predictions within the relevant region, reporting the ten largest $\Delta_t^{(y)}$ (by absolute value) for $y \in \{0,1\}$ and $\lambda \in \{0,3\}$, computed on the $|s(X) - \tau| \le 0.5$ band. We observe that collateral and program/product variables consistently contribute at high to moderate magnitudes and with stable signs across outcome groups and fairness regimes. `loan_purpose_1` (home purchase) is the single

[20] Specifically, we focus on the window $|s_\theta(X) - \tau| \le 0.5$ (i.e., predicted probabilities in $[0.38, 0.62]$), explicitly to stay in the locally-linear, decision-relevant region.

largest negative contributor near the threshold for both classes and both λ's (from -0.200 to -0.153 in denials and from -0.298 to -0.258 in originations), implying that, conditional on Y, purchase purpose lowers the average score for minority applicants relative to non-minorities. The univariate `loan_to_value_ratio` term is positive and rises slightly with fairness regularization (from 0.039 to 0.040 in denials and from 0.052 to 0.057 in originations), indicating that near-threshold differences in LTV continue to explain part of the gap. The interaction `log1p(loan_amount)`×`log1p(property_value)` is positive across λ and Y, consistent with an exposure-collateral channel at the margin. The interaction `dti_num`×`loan_type_1` (conventional) is negative and stable (about -0.045 for Y=1 and -0.06 for Y=0), suggesting a gap-reducing DTI impact dependent on insured/guaranteed loan status near the operating threshold. Age composition terms are consistently small and negative (e.g., `applicant_age_25-34` around -0.012) and do not drive the gap.

Moreover, repayment capacity remains a primary driver at the margin. In denials (Y=1), `dti_num` is large and positive at both λ=0 and λ=3 (0.111 and 0.105), indicating that, among applications near the decision boundary that are denied, minority applicants carry higher DTI-induced scores on average. In originations (Y=0), `dti_num` is again positive and slightly larger under fairness (0.167 at λ=0, 0.174 at λ=3), showing that capacity continues to drive approvals near the threshold.

In contrast, geographic proxies compress under fairness regularization. `tract_minority_population_percent` is sizable in both outcome subsets at λ=0 (0.131 for Y=1 and 0.147 for Y=0) and shrinks considerably when penalization is enforced (to 0.038 for Y=1 and 0.041 for Y=0), consistent with a reweighting away from proxy-like variables at the operating margin. Other measures such as `tract_to_msa_income_percentage` or `tract_population` remain small and stable (about 0.016-0.020 across outcomes and λ), indicating that the reduction is not simply a reallocation to other geographic variables.

Overall, the fairness penalty reduces dependence on geographic correlates while preserving economically grounded channels, such as repayment capacity and collateral, at moderate magnitudes.[21] The residual disparities near the threshold are dispersed across core underwriting determinants, indicating improved parity without flattening fundamental risk signals.[22]

[21] That is increasing λ reduces EO disparity by reweighting away from proxy-like geographic correlates at the decision margin, while leaving economically grounded risk signals (capacity/collateral) in place.

[22] As we show in Appendix M, similar conclusions follow from a decomposition of the score gap on the full data.

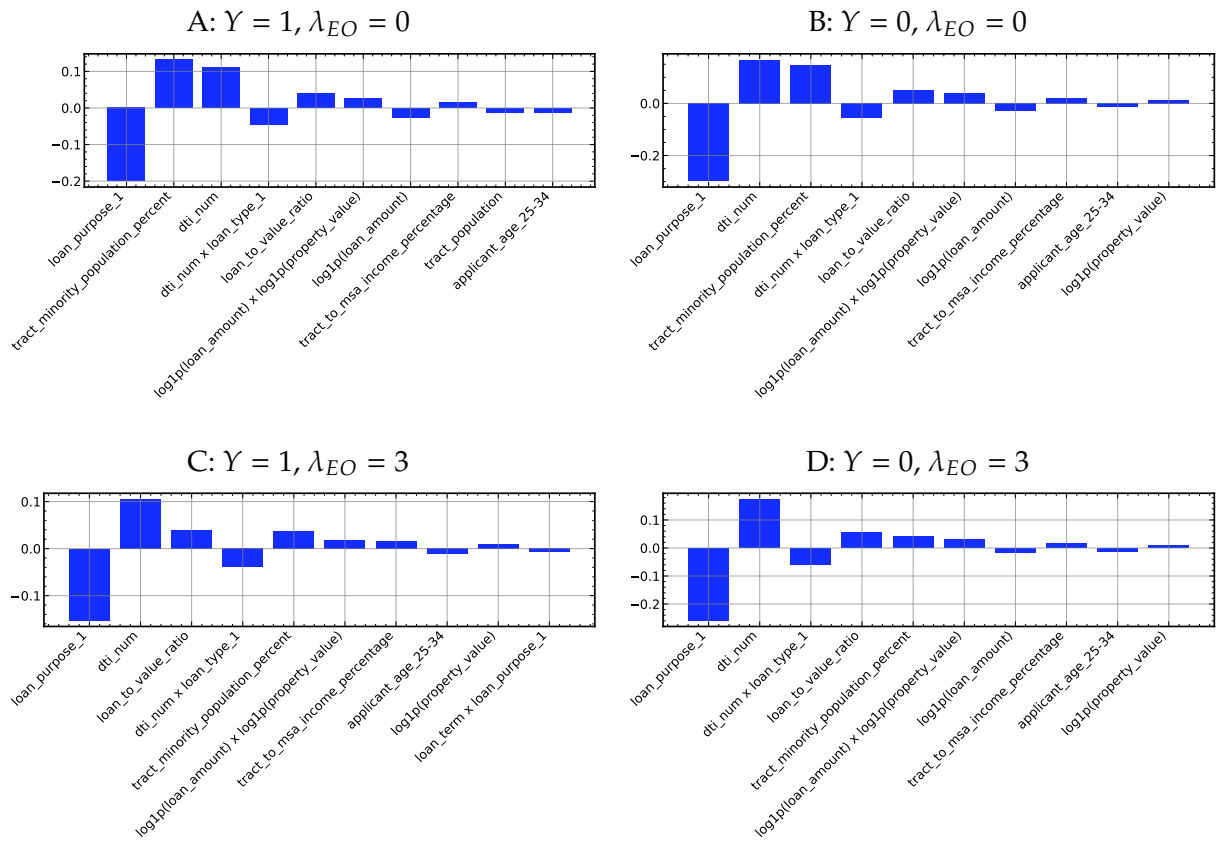


Figure 7: $\Delta_t^{(y)}$ across class labels and models. This figure compares the most important contributors to the EO gap for different outcomes and penalization levels. For each case, the ten largest $\Delta_t^{(y)}$ (by absolute value) are reported. The decomposition of the mean model score gap into term-wise components $\Delta_t^{(y)}$ explains predicted probability gaps around the decision threshold.

5.3 Within-lender fairness

In practice, loan approval is not determined by a single decision maker, but by many lenders, each operating with its own risk assessment and modeling workflow. This raises a natural question: are the EO/fairness gains coming from changing who gets approved *within* a given lender (i.e., improving each lender's internal ranking/decision rule), or are they driven by reallocation *across* lenders (e.g., because minority applicants sort into different lenders, or because the model implicitly shifts volume across institutions)? In this section, we address these questions by examining whether the fairness gains persist lender-by-lender and decomposing the overall disparity into within- and between-lender components.

To this end, we hold each lender's denial rate fixed and ask whether the model improves the within-lender ordering of applicants, rather than merely reallocating approvals and denials across institutions. Specifically, we convert scores into counterfactual model decisions by choosing, for each lender ℓ, a score threshold that exactly reproduces that lender's observed denial rate on the test set. This "capacity matching" fixes overall approval volume at each lender, so any change in group disparities cannot come from mechanically shifting approvals across institutions. We then calculate $\Delta\text{TPR}_\ell = \text{TPR}_\ell^{(1)} - \text{TPR}_\ell^{(0)}$ and $\Delta\text{FPR}_\ell = \text{FPR}_\ell^{(1)} - \text{FPR}_\ell^{(0)}$ at the lender level and summarize the cross-lender distribution of these gaps by volume-weighted means and standard deviations, using each lender's test-set volume as weights. To isolate where disparities originate, we apply an Oaxaca-Blinder-style decomposition (Blinder, 1973; Oaxaca, 1973), splitting the pooled gap into a within-lender component and a residual between-lender component: As above, let $A \in \{0, 1\}$ denote the protected-group indicator (e.g., $A = 1$ minority, $A = 0$ non-minority) and, further, let L be the lender identifier taking values $\ell \in \mathcal{L}$ and let $Y_{\text{obs}} \in \{0, 1\}$ be the observed outcome and $Y_m \in \{0, 1\}$ be the model's approve/deny decision. For each lender ℓ and group a, we define the within-lender approval rate in the chosen subset as

$$p_{\ell a}(y) \;=\; \Pr\big(Y_m = 1 \mid L = \ell,\; A = a,\; Y_{\text{obs}} = y\big)\,.$$

Additionally, we define the group-specific lender share in that segment as

$$s_{\ell a}(y) \;=\; \Pr\big(L = \ell \mid A = a,\; Y_{\text{obs}} = y\big)\,, \qquad \sum_{\ell \in \mathcal{L}} s_{\ell a}(y) = 1,$$

and the pooled (group-agnostic) lender share as

$$w_\ell(y) \;=\; \Pr\big(L=\ell \mid Y_{\text{obs}}=y\big) \;=\; \sum_{a\in\{0,1\}} \Pr\big(A=a \mid Y_{\text{obs}}=y\big)\; s_{\ell a}(y), \qquad \sum_{\ell\in\mathcal{L}} w_\ell(y)=1.$$

We are interested in the pooled (total) gap in the outcome group y, defined as the difference of the groups' approval probabilities

$$\Delta_{\text{total}}(y) \;=\; \Pr\big(Y_m=1 \mid A=1,\; Y_{\text{obs}}=y\big) - \Pr\big(Y_m=1 \mid A=0,\; Y_{\text{obs}}=y\big) \;= \\ \sum_{\ell\in\mathcal{L}} s_{\ell 1}(y)\, p_{\ell 1}(y) \;-\; \sum_{\ell\in\mathcal{L}} s_{\ell 0}(y)\, p_{\ell 0}(y).$$

Note that fixing $y=1$ corresponds to the true-positive subset (TPR analysis) and $y=0$ to the false-positive subset (FPR analysis). Adding and subtracting the same terms $\sum_\ell w_\ell(y)\, p_{\ell 1}(y)$ and $\sum_\ell w_\ell(y)\, p_{\ell 0}(y)$ yields the exact decomposition

$$\Delta_{\text{total}}(y) \;=\; \underbrace{\sum_{\ell\in\mathcal{L}} w_\ell(y)\,\big[p_{\ell 1}(y)-p_{\ell 0}(y)\big]}_{\Delta_{\text{within}}(y)} \;+ \\ \underbrace{\sum_{\ell\in\mathcal{L}} \big[s_{\ell 1}(y)-w_\ell(y)\big]\, p_{\ell 1}(y) \;-\; \sum_{\ell\in\mathcal{L}} \big[s_{\ell 0}(y)-w_\ell(y)\big]\, p_{\ell 0}(y)}_{\Delta_{\text{between}}(y)} \tag{9}$$

into a within- and a between-lender component. The within term is a lender-level average of per-lender gaps $p_{\ell 1}(y)-p_{\ell 0}(y)$ taken with a common set of weights $w_\ell(y)$ that do not depend on group membership. It therefore answers the counterfactual "what gap would arise if both groups faced the same lender mix in this subset while keeping each lender's behavior unchanged?". It represents a pure average of within-lender differences, so the only way it moves is if the model treats the two groups differently within the application sets associated with individual lenders. The between term is the remainder that depends on differences in lender composition across groups, through $s_{\ell a}(y)-w_\ell(y)$, interacting with lender-specific baseline approval levels $p_{\ell a}(y)$. This fomulation directly implies two special cases: if groups have identical lender mixes $s_{\ell 1}(y)=s_{\ell 0}(y)=w_\ell(y)$ for all ℓ, then $\Delta_{\text{between}}(y)=0$ and the pooled gap equals the within component. If per-lender behavior is group-invariant $p_{\ell 1}(y)=p_{\ell 0}(y)$ for all ℓ, then $\Delta_{\text{within}}(y)=0$ and any pooled gap must come from composition differences.

Table 4 reports the results for the baseline model (FAN, λ=0) and fairness-penalized

variants ($\lambda \in \{1, 2, 3\}$).[23] With capacity held fixed for each lender, the baseline exhibits a positive cross-lender weighted mean ΔTPR_ℓ of 0.08 and ΔFPR_ℓ of 0.042, indicating that, conditional on the observed outcome, the model is more likely to reject minority applicants than otherwise-similar non-minority applicants both among actually denied loans (true positives) and among actually approved loans (false positives), which is in line with our previous findings. The within-between decomposition shows that these gaps are almost entirely *within* lenders (e.g., ΔTPR: total 0.0756, within 0.0757, between -0.0001), implying that disparities are generated by the model's intra-lender scoring process rather than by differences in which lenders minority borrowers approach.

Introducing the fairness penalty compresses these within-lender gaps monotonically across λ. At λ=3, the weighted mean ΔTPR_ℓ falls to 0.041 while mean ΔFPR_ℓ falls to 0.014, with the between-lender terms remaining negligible. Economically, this indicates that fairness regularization reshapes the local trade-offs the model makes among applicants associated with an individual lender, e.g., along DTI/LTV and purpose/type margins, while each lender's overall denial rate is preserved. Because we match capacities at the lender level, any change in pooled disparities cannot come from shifting approval volumes across institutions. As the negligible between-lender component indicates, disparities are also not driven by differential sorting of groups across lenders with systematically different conditional error rates under our model. Rather, consistently across fairness penalization levels, the gap arises almost entirely from the model's within-lender ordering near the decision boundary. This suggests that, in practice, lenders could meaningfully reduce disparities by modifying their internal scoring rules, even without changing pricing, capacity, or market structure.

5.4 Risk parity and pricing alignment

Finally, we examine whether fairness regularization reduces cross-group disparities among observed approvals (Y=0) where pricing information is available and whether those fairness gains leave pricing relationships intact for approved loans. In consumer credit, supervisory and economic concerns jointly require that (i) applicants near the approval cutoff are treated comparably across protected groups and (ii) conditional on modeled risk within approvals, pricing does not develop group-specific distortions. Therefore, we report group means of the model's rejection signal among approved loans and, separately,

[23] The "Mean" $\Delta\text{TPR}_\ell/\Delta\text{FPR}_\ell$ columns average lender-level gaps using each lender's test-set volume as weights, whereas the "Total" ΔTPR/ΔFPR columns report the single pooled gap on the full test set within the corresponding outcome group. These differ in general because they use different weighting schemes.

		ΔTPR_ℓ		ΔFPR_ℓ		ΔTPR Decomp.			ΔFPR Decomp.		
Model	**N**	Mean	W-Std	Mean	W-Std	Total	Within	Between	Total	Within	Between
FAN ($\lambda = 0$)	45	0.0804	0.0576	0.0419	0.0254	0.0756	0.0757	-0.0001	0.0431	0.0433	-0.0002
FAN ($\lambda = 1$)	45	0.0647	0.0602	0.0298	0.0242	0.0615	0.0616	-0.0001	0.0299	0.0301	-0.0001
FAN ($\lambda = 2$)	45	0.0519	0.0636	0.0229	0.0218	0.0476	0.0488	-0.0012	0.0229	0.0230	-0.0001
FAN ($\lambda = 3$)	45	0.0408	0.0595	0.0142	0.0235	0.0395	0.0400	-0.0005	0.0144	0.0147	-0.0003

Table 4: Within-lender, capacity-matched fairness by model. This table reports lender-level TPR and FPR gaps for different levels of fairness regularization. N is the number of lenders included in the analysis (i.e., those with sufficient test-set observations in each outcome-by-group cell to reliably compute rates). The ΔTPR_ℓ / ΔFPR_ℓ columns summarize average lender-level gaps using each lender's test-set volume as weights, whereas the "Total" ΔTPR/ΔFPR columns use the single pooled gap on the full test set within the corresponding outcome group.

study risk-price alignment on the approved loans.

Table 5 reports group averages of the predicted rejection probability $p = \hat{p}(X)$ and the raw model score $s_\theta(X)$ among Y=0, together with cross-group gaps Δ defined as the difference between the minority and the non-minority group. For the unconstrained classifier (λ=0), approved minority applicants exhibit higher average modeled rejection risk than approved non-minority applicants ($\Delta\overline{p}$ = 2.32 p.p., $\Delta\overline{s_\theta}$ = 0.191), indicating residual disparity at the approval frontier. As the fairness weight increases, these gaps shrink materially: $\Delta\overline{p}$ declines to 1.72 p.p. at λ=1, 1.31 p.p. at λ=2, and 1.14 p.p. at λ=3. On the score scale, which avoids the compression of percentage-point probability differences, the reduction is especially clear, falling from 0.191 at λ=0 to 0.040 at λ=2 and settling at 0.107 at λ=3.

Table 6 evaluates pricing alignment within the approved set for two observed HMDA price measures, interest rate and rate spread. For each model, we report (i) the correlation between s_θ and the pricing variable by group and (ii) the conditional minority coefficient $\hat{\gamma}$ from the pooled regression

$$\text{price}_l = \alpha + \beta\, s_\theta(X_l) + \gamma\, \mathbf{1}\{A_l{=}1\} + \varepsilon_l \tag{10}$$

estimated on approvals. We find that the risk-price correlations are near zero across groups and across λ values. Crucially, the conditional minority coefficient remains both statistically and economically (in magnitude) insignificant as we raise the fairness weight: for interest rate, $\hat{\gamma} \approx$ 0.0065-0.0068 percentage points (roughly 0.65-0.68 bps), and for rate spread, $\hat{\gamma} \approx$ 0.011-0.012 percentage points (about 1.1-1.2 bps). We therefore find no evidence that fairness regularization induces pricing asymmetries or cross-subsidization

within approved loans.

Taken together, these results demonstrate that the penalized classifier reduces group disparities in the model's rejection signal precisely at the approval frontier while leaving pricing relationships on the approved loans essentially unchanged after conditioning on modeled risk. The approach, therefore, achieves its intended parity improvements in the decision region that matters in practice without distorting observed pricing relationships among the loans that are actually originated.

	$\overline{p}_0$	$\overline{p}_1$	$\Delta\overline{p}$	$\overline{s_\theta(X)}_0$	$\overline{s_\theta(X)}_1$	$\Delta\overline{s_\theta(X)}$	$N(Y{=}0)$	N_0	N_1
FAN ($\lambda = 0$)	0.1546	0.1778	0.0232	-1.9739	-1.7829	0.1910	57873	46241	11632
FAN ($\lambda = 1$)	0.1562	0.1734	0.0172	-1.9585	-1.8124	0.1460	57873	46241	11632
FAN ($\lambda = 2$)	0.1515	0.1646	0.0131	-2.3465	-2.3063	0.0402	57873	46241	11632
FAN ($\lambda = 3$)	0.1605	0.1720	0.0114	-1.9130	-1.8065	0.1065	57873	46241	11632

Table 5: Risk Parity by Model (Y=0). This table compares model outputs on approved loans for the minority and the non-minority group and across different levels of fairness regularization.

Model	N	Corr$(s_\theta(X), \cdot)_0$	Corr$(s_\theta(X), \cdot)_1$	$\hat{\gamma}$ (minority)
Panel A: Interest rate				
FAN ($\lambda = 0$)	39264	-0.0022	-0.0008	0.0068
FAN ($\lambda = 1$)	39264	-0.0019	-0.0002	0.0067
FAN ($\lambda = 2$)	39264	-0.0012	-0.0014	0.0065
FAN ($\lambda = 3$)	39264	-0.0018	0.0009	0.0066
Panel B: Rate spread				
FAN ($\lambda = 0$)	32800	0.0058	0.0019	0.0111
FAN ($\lambda = 1$)	32800	0.0061	0.0034	0.0113
FAN ($\lambda = 2$)	32800	0.0012	0.0098	0.0119
FAN ($\lambda = 3$)	32800	0.0063	0.0029	0.0114

Table 6: Pricing alignment (Y=0). This table reports correlations between model scores and pricing variables as well as minority differentials estimated from specification (10). We consider the two HMDA pricing measures interest rate (Panel A) and rate spread (Panel B).

6 Conclusions

We develop an interpretable fairness-aware machine learning classifier that combines predictive accuracy and fairness constraints in the objective function. In the public-HMDA setting, this allows us to construct alternative decision rules that generate less disparate predicted denial and origination outcomes across minority groups while maintaining strong out-of-sample fit to historical underwriting labels. Leveraging its interpretable structure, we show the impact of individual input variables on predictions and examine how the prediction rules are adjusted to account for the fairness objective, complementing regulatory requirements. We find that lenders heavily weight loan-to-value and debt-to-income ratios, penalize very low or very high loan amounts, and treat non-purchase loans as riskier, possibly due to higher leverage and weaker collateral signals.

Additional analyses provide insights into the underlying economic mechanisms: Because the output decomposes into feature contributions, we can identify which variables drive denials overall and for specific applications, translate these drivers into adverse action reasons, and monitor how fairness mitigation changes this composition. Feature-level decompositions of the equalized-odds gap show that fairness regularization reduces the contribution of geographic proxies and concentrates the decision on core underwriting factors such as debt-to-income. Lender-level results indicate that nearly all observed disparities arise from the models' ordering of applicants within institutions rather than from differences in lender choice, and that the fairness penalty compresses these gaps monotonically. Among other robustness tests, we demonstrate that the fairness adjustments persist across time, decision thresholds, and applicant subgroups while preserving the main pricing relationships observed in the data.

Taken together, our results show that within a public-HMDA information set, lenders and researchers can construct interpretable, fairness-constrained scoring rules that reduce label-conditional disparity measures with negligible loss in predictive fit to historical underwriting outcomes.

These findings have two implications. First, from a fair-lending perspective, the existence of such models means that status-quo scoring rules are unlikely to be "least discriminatory", strengthening the case for LDA-style regulation and supervisory guidance that explicitly requires a search over fairness-aware alternatives. Second, from a research perspective, the relatively flat fit-fairness trade-off in this application suggests that fairness constraints need not substantially degrade a model's ability to approximate historical underwriting outcomes, though richer private data would be required to assess implications for welfare, profitability, and fairness conditional on true borrower risk. To this end, our

framework can be applied to other products and outcomes such as pricing, default, or refinancing to map out similar frontiers in other areas of consumer finance.

Appendices

A Summary statistics

	Count	Mean	S.D.	25%	50%	75%
dti_num	309187	46.893	9.002	40.000	45.000	49.000
ffiec_msa_md_median_family_income	500000	74184.045	16609.678	64200.000	73500.000	81400.000
income	489532	96.371	162.949	49.000	73.000	113.000
loan_amount	500000	244933.600	272797.520	125000.000	205000.000	305000.000
loan_term	496697	337.036	59.856	360.000	360.000	360.000
loan_to_value_ratio	472617	86.896	1900.862	69.650	80.000	96.499
property_value	490292	340662.687	539492.861	165000.000	255000.000	395000.000
tract_median_age_of_housing_units	500000	34.357	17.496	21.000	33.000	45.000
tract_minority_population_percent	500000	31.799	25.347	11.680	24.310	46.220
tract_one_to_four_family_homes	500000	2007.548	1113.764	1347.000	1827.000	2435.000
tract_owner_occupied_units	500000	1487.588	896.788	950.000	1345.000	1833.000
tract_population	500000	5753.300	3256.018	3874.000	5195.000	6831.000
tract_to_msa_income_percentage	500000	112.943	41.763	87.000	108.000	134.000

Table 7: Summary statistics of raw data (continuous predictors)

Variable	Level	Count (pct)
applicant_age	25-34	107625 (21.5%)
	35-44	111859 (22.4%)
	45-54	104120 (20.8%)
	55-64	84641 (16.9%)
	65-74	54587 (10.9%)
	Other	37168 (7.4%)
applicant_sex	1	328423 (65.7%)
	2	169810 (34.0%)
	3	1447 (0.3%)
	4	2 (0.0%)
	6	318 (0.1%)
business_or_commercial_purpose	1	304 (0.1%)
	1111	79 (0.0%)
	2	499617 (99.9%)
conforming_loan_limit	C	476660 (95.3%)
	NC	23303 (4.7%)
	U	37 (0.0%)
construction_method	1	481534 (96.3%)
	2	18466 (3.7%)
derived_dwelling_category	Single Family (1-4 Units):Manufactured	18466 (3.7%)
	Single Family (1-4 Units):Site-Built	481534 (96.3%)
dti_missing	0	309187 (61.8%)
	1	190813 (38.2%)
loan_purpose	1	281568 (56.3%)
	2	14544 (2.9%)
	31	73959 (14.8%)
	32	115156 (23.0%)
	4	14504 (2.9%)

Continued on next page

Variable	Level	Count (pct)
	Other	269 (0.1%)
loan_type	1	339116 (67.8%)
	2	93180 (18.6%)
	3	60239 (12.0%)
	4	7465 (1.5%)
manufactured_home_land_property_interest	1	12903 (2.6%)
	2	419 (0.1%)
	3	3691 (0.7%)
	4	1412 (0.3%)
	5	481496 (96.3%)
	Other	79 (0.0%)
manufactured_home_secured_property_type	1	10887 (2.2%)
	1111	79 (0.0%)
	2	7556 (1.5%)
	3	481478 (96.3%)
open-end_line_of_credit	1	35548 (7.1%)
	1111	79 (0.0%)
	2	464373 (92.9%)
preapproval	1	14987 (3.0%)
	2	485013 (97.0%)
reverse_mortgage	1	2460 (0.5%)
	1111	79 (0.0%)
	2	497461 (99.5%)
state_code	CA	57069 (11.4%)
	FL	42361 (8.5%)
	NC	20366 (4.1%)
	OH	21525 (4.3%)
	TX	38939 (7.8%)
	Other	318711 (63.9%)

Table 8: This table presents value counts for categorical predictors in our data set. For variables with many levels, the top categories are displayed and the remainder are grouped as "Other".

B Variable definitions

Abbreviation	HMDA Variable	Values / HMDA Codes
applicant_age	Applicant Age	<25 – Age is less than 25; 25-34 – Age is between 25 and 34; 35-44 – Age is between 35 and 44; 45-54 – Age is between 45 and 54; 55-64 – Age is between 55 and 64; 65-74 – Age is between 65 and 74; >74 – Age is greater than 74; 8888 – Not applicable
applicant_sex	Applicant Sex	1 – Male; 2 – Female; 3 – Information not provided by applicant in mail, internet, or telephone application; 4 – Not applicable; 6 – Applicant selected both male and female
business_or_commercial_purpose	Business or Commercial Purpose	1 – Primarily for a business or commercial purpose; 2 – Not primarily for a business or commercial purpose; 1111 – Exempt
conforming_loan_limit	Conforming Loan Limit	C – Conforming; NC – Nonconforming; U – Undetermined; NA - Not Applicable
construction_method	Loan Purpose	1 – Home purchase; 2 – Home improvement; 3 – Refinancing; 31 – Cash-out refinancing; 32 – Other purpose; 4 – Not applicable.
ddc	Derived Dwelling Category	sf_site_built – Single Family (1-4 Units):Site-Built; mf_site_built – Multifamily:Site-Built (5+ Units); sf_manufactured – Single Family (1-4 Units):Manufactured; mf_manufactured – Multifamily:Manufactured (5+ Units)
dti_missing	Debt-to-Income Ratio	0 – Available debt-to-income information; 1 – Missing debt-to-income information
dti_num	Debt-to-Income Ratio	Varying values

Continued on next page

Abbreviation	HMDA Variable	Values / HMDA Codes
ffiec_msa_md_median_family_income	FFIEC Median family income of the MSA/MD[24]	Varying values
income	Income	Varying values
loan_amount	Loan Amount	Varying values
loan_purpose	Loan Purpose	1 – Home purchase; 2 – Home improvement; 3 – Refinancing; 31 – Cash-out refinancing; 32 – Other purpose; 4 – Not applicable.
loan_term	Loan Term	Varying values
loan_to_value_ratio	Loan-to-value Ratio	Varying values
loan_type	Loan Type	1 – Conventional (not insured or guaranteed by FHA VA RHS or FSA); 2 – Federal Housing Administration insured (FHA); 3 – Veterans Affairs guranteed (VA); 4 – USDA Rural Housing Service or Farm Service Agency guaranteed (RHS or FSA)
mfd_home_lp_interest	Manufactured Home Land Property Interest	1 – Direct ownership; 2 – Indirect ownership; 3 – Paid leasehold; 4 – Unpaid leasehold; 5 – Not applicable; 1111 - Exempt
mfd_home_sec_prop_type	Manufactured Home Secured Property Type	1 – Manufactured home and land; 2 – Manufactured home and not land; 3 – Not applicable; 1111 – Exempt.
open-end_line_of_credit	Open-End Line of Credit	1 – Open-end line of credit; 2 – Not an open-end line of credit; 1111 – Exempt
preapproval	Preapproval	1 – Preapproval requested; 2 – Preapproval not requested
property_value	Property Value	Varying values

Continued on next page

[24]Median family income for a Metropolitan Statistical Area (MSA) as published by the Federal Financial Institutions Examination Council.

Abbreviation	HMDA Variable	Values / HMDA Codes
reverse_mortgage	Reverse Mortgage	1 – Reverse mortgage; 2 – Not a reverse mortgage; 1111 – Exempt
state_code	State Code	AL – Alabama; ...; WY – Wyoming
tract_median_age_of_housing_units	Tract Median Age of Housing Units	Varying values
tract_minority_population_percent	Tract Minority Population Percent	Varying values
tract_one_to_four_family_homes	Tract 1-4 Family Homes	Varying values
tract_owner_occupied_units	Tract Owner-occupied Units	Varying values
tract_population	Tract Population	Varying values
tract_to_msa_income_percentage	Tract-to-MSA Income Percentage	Varying values

Table 9: This table contains variable definitions and descriptions for our data set.

C Model: lender profit maximization

In the following, we formalize how the binary cross-entropy (BCE) loss augmented with a fairness regularizer arises from a simple behavioral model of a risk-neutral lender that maximizes expected profit while being subject to (or voluntarily adopting) a penalty on disparate decisions. Note that this represents a stylized conceptual model rather than an identified empirical model of the public-HMDA setting. Public HMDA does not contain the repayment outcomes, lender payoffs, credit scores, or other private underwriting information needed to estimate the latent objects in this framework, so the model should be read only as motivation for the empirical specification. We start from the assumption that the lender's operational decision is whether to approve a loan, that profits accrue only when a loan is granted, and that the only source of uncertainty is repayment. We then (i) derive the profit-maximizing decision rule as a threshold on the true repayment probability, (ii) explain the role of false positives and false negatives by decomposing expected profit into confusion-matrix terms with unequal costs, (iii) justify BCE minimization as a strictly proper surrogate for learning the profit-relevant repayment probabilities, and (iv) obtain our training objective as a Lagrangian relaxation in which BCE learns the probabilities needed for profit maximization and the fairness term controls disparities in approval or error rates.

Applicants are described by features $X \in \mathcal{X}$ and a protected attribute $A \in \mathcal{A}$ (e.g., $A \in \{0,1\}$). The repayment outcome is $Y \in \{0,1\}$, where $Y = 1$ denotes full repayment. If a loan is granted, the lender earns a net return $R > 0$ when $Y = 1$ and incurs a loss $C > 0$ when $Y = 0$. If the loan is denied, profit is 0. Let the true conditional repayment probability be

$$p^\star(x,a) \;=\; \mathbb{P}(Y = 1 \mid X = x, A = a).$$

A (possibly data-driven) approval rule $d : \mathcal{X} \times \mathcal{A} \to \{0,1\}$ delivers pointwise expected profit

$$\pi(x,a;d) \;=\; d(x,a)\Big[p^\star(x,a)\,R \;-\; \big(1 - p^\star(x,a)\big)\,C\Big].$$

Thus the Bayes-optimal rule approves if and only if the expected gain from approval is nonnegative, i.e.

$$d^\star(x,a) \;=\; \mathbb{I}\Big\{\, p^\star(x,a) \;\geq\; t^\star \,\Big\}, \qquad t^\star \;=\; \frac{C}{R+C}.$$

With heterogeneous offers (R_i, C_i) across applicants i, the cutoff becomes $t_i^\star = C_i/(R_i + C_i)$.

It is useful to connect this threshold rule to the false positive/false negative (FP/FN)

logic. Let $d = 1$ denote approval (a positive decision). A *true positive* (TP) is $(d = 1, Y = 1)$, a *false positive* (FP) is $(d = 1, Y = 0)$, a *false negative* (FN) is $(d = 0, Y = 1)$, and a *true negative* (TN) is $(d = 0, Y = 0)$. Aggregating over the population, expected profit can be written as

$$\Pi(d) \;=\; \mathbb{E}\big[\pi(X, A; d)\big] \;=\; R\,\mathbb{P}(d = 1, Y = 1) \;-\; C\,\mathbb{P}(d = 1, Y = 0) \;=\; R \cdot \mathrm{TP} \;-\; C \cdot \mathrm{FP}.$$

Therefore, FPs destroy C units of profit each, TPs create R units each, TNs contribute 0, and FNs matter via forgone TPs (each FN forgoes a potential R). Equivalently, profit maximization is the same as minimizing the expected cost-sensitive misclassification loss with costs

$$\mathrm{cost}(\mathrm{FP}) = C, \qquad \mathrm{cost}(\mathrm{FN}) = R,$$

because approving yields a cost only when $Y = 0$ (size C), while denying yields a cost only when $Y = 1$ (size R). The Bayes rule for this asymmetric 0–1 loss is precisely to predict 1 when $p^\star(x, a) \geq C/(R{+}C)$.

The FP/FN view also clarifies how groupwise error rates interact with profit. Writing $\pi_a = \mathbb{P}(A = a)$, the groupwise true- and false-positive rates

$$\mathrm{TPR}_a(d) \;=\; \mathbb{P}(d = 1 \mid Y = 1, A = a), \qquad \mathrm{FPR}_a(d) \;=\; \mathbb{P}(d = 1 \mid Y = 0, A = a),$$

yield the decomposition

$$\Pi(d) \;=\; R \sum_a \pi_a\, \mathbb{P}(Y = 1 \mid A = a)\, \mathrm{TPR}_a(d) \;-\; C \sum_a \pi_a\, \mathbb{P}(Y = 0 \mid A = a)\, \mathrm{FPR}_a(d).$$

Many fairness definitions (e.g., demographic parity, equal opportunity, equalized odds) are naturally expressed in terms of approval rates and these same error rates.

In practice, $p^\star$ is unknown and must be learned. Let a parametric model $\hat{p}_\theta(x, a) \in (0, 1)$ approximate $p^\star(x, a)$. Since the optimal decision depends only on $p^\star$ via a threshold at $t^\star$, we should estimate $p^\star$ as accurately and calibratedly as possible. A standard approach is to minimize a strictly proper scoring rule for Bernoulli outcome, for example the log-loss (binary cross-entropy),

$$\ell_{\mathrm{BCE}}(\theta) \;=\; \mathbb{E}\big[-Y \log \hat{p}_\theta(X, A) \;-\; (1 - Y) \log\big(1 - \hat{p}_\theta(X, A)\big)\big]\,.$$

Because

$$\mathbb{E}\left[-Y\log q(X,A) \;-\; (1-Y)\log\big(1-q(X,A)\big)\right] \;=\; \underbrace{\mathbb{E}\big[H\big(p^\star(X,A)\big)\big]}_{\text{irreducible}} \;+\; \mathbb{E}\Big[\mathrm{KL}\big(\mathrm{Bern}(p^\star(X,A))\,\|\,\mathrm{Bern}(q(X,A))\big)\Big],$$

the unique population minimizer is $q = \hat{p}_\theta = p^\star$.[25] Empirically, given data $\{(x_i, a_i, y_i)\}_{i=1}^n$,

$$\widehat{\ell}_{\mathrm{BCE}}(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big(-y_i\log\hat{p}_\theta(x_i,a_i) \;-\; (1-y_i)\log\big(1-\hat{p}_\theta(x_i,a_i)\big)\Big).$$

Minimizing $\widehat{\ell}_{\mathrm{BCE}}$ yields a consistent estimator of $p^\star$ under standard conditions. Approving by thresholding

$$d_\theta(x,a) \;=\; \mathbb{I}\big\{\hat{p}_\theta(x,a) \geq t^\star\big\}$$

then converges to the Bayes profit-maximizing rule as $\hat{p}_\theta \to p^\star$. In this sense, minimizing BCE is a surrogate for maximizing expected profit: BCE learns the probabilities that, when thresholded at $t^\star = C/(R{+}C)$ (or $t_i^\star = C_i/(R_i{+}C_i)$ in the heterogeneous case), implement the optimal cost-sensitive classifier with cost(FP)=C and cost(FN)=R.[26]

Fairness considerations enter either as hard constraints on the induced policy or as a penalty that charges for unfairness. Let $J_{\mathrm{fair}}(\theta)$ measure unfairness for the policy induced by $\hat{p}_\theta$. Examples include demographic parity (DP),

$$J_{\mathrm{fair}}^{\mathrm{DP}}(\theta) \;=\; \Big(\mathbb{E}\big[d_\theta(X,A)\mid A=1\big] - \mathbb{E}\big[d_\theta(X,A)\mid A=0\big]\Big)^2,$$

and equal opportunity (EqO),

$$J_{\mathrm{fair}}^{\mathrm{EqO}}(\theta) \;=\; \Big(\mathbb{E}\big[d_\theta(X,A)\mid A=1, Y=1\big] - \mathbb{E}\big[d_\theta(X,A)\mid A=0, Y=1\big]\Big)^2,$$

with equalized odds obtained by combining EqO with a similar penalty on false-positive rates. Because d_θ contains a non-differentiable indicator, we compute a smooth surrogate

[25] Equivalently, one can verify pointwise optimality by differentiation: for fixed $p^\star \in (0,1)$,

$$\frac{\partial}{\partial q}\{-p^\star\log q - (1-p^\star)\log(1-q)\} = -\tfrac{p^\star}{q} + \tfrac{1-p^\star}{1-q} = 0$$

implies $q = p^\star$, and the second derivative $\frac{p^\star}{q^2} + \frac{1-p^\star}{(1-q)^2} > 0$ ensures a unique minimum.

[26] Directly maximizing empirical profit over θ is difficult because the indicator $\mathbb{I}\{\hat{p}_\theta \geq t^\star\}$ makes the objective nonsmooth and nonconvex. BCE avoids this by learning $p^\star$ via a strictly proper, smooth surrogate.

during training by replacing d_θ with

$$s_\theta(x,a) \;=\; \sigma\!\left(\tfrac{\hat{p}_\theta(x,a)-t^\star}{\tau}\right), \qquad \tau>0,$$

where σ is the logistic sigmoid and τ is a temperature, with the property that as $\tau \to 0$, $s_\theta \to d_\theta$ pointwise. In terms of error-rate surrogates, we can write

$$\widetilde{\mathrm{TPR}}_a(\theta) \;=\; \mathbb{E}\big[s_\theta(X,A) \mid Y=1, A=a\big], \qquad \widetilde{\mathrm{FPR}}_a(\theta) \;=\; \mathbb{E}\big[s_\theta(X,A) \mid Y=0, A=a\big],$$

and define, for instance, an equalized-odds penalty

$$J_{\mathrm{fair}}^{\mathrm{EOdds}}(\theta) \;=\; \Big(\widetilde{\mathrm{TPR}}_1(\theta)-\widetilde{\mathrm{TPR}}_0(\theta)\Big)^2 \;+\; \Big(\widetilde{\mathrm{FPR}}_1(\theta)-\widetilde{\mathrm{FPR}}_0(\theta)\Big)^2.$$

A regulator- or lender-preferred formulation is the constrained problem

$$\max_\theta \;\; \Pi(\theta) \quad \text{s.t.} \quad G(\theta) \;\le\; \varepsilon, \qquad \Pi(\theta)=\mathbb{E}\Big[d_\theta(X,A)\Big(p^\star(X,A)R-\big(1-p^\star(X,A)\big)C\Big)\Big],$$

where $G(\theta)$ is a chosen unfairness functional (e.g., $G=J_{\mathrm{fair}}^{1/2}$) and $\varepsilon \ge 0$ is a tolerance. The corresponding Lagrangian is

$$\mathcal{L}_{\mathrm{lag}}(\theta,\lambda) \;=\; \Pi(\theta) \;-\; \lambda\, G(\theta), \qquad \lambda \ge 0.$$

Optimizing $\Pi(\theta)$ directly is intractable due to the discontinuity of d_θ. Replacing the non-smooth profit term by the strictly proper surrogate that learns $p^\star$ yields the tractable empirical objective

$$\widehat{\mathcal{L}}(\theta;\lambda) \;=\; \widehat{\ell}_{\mathrm{BCE}}(\theta) \;+\; \lambda\, J_{\mathrm{fair}}(\theta),$$

which is a Lagrangian relaxation of profit maximization under fairness. The BCE term learns the calibrated repayment probabilities that are sufficient for profit maximization when paired with the threshold at $t^\star = C/(R{+}C)$ (or loan-specific $t_i^\star$), while $J_{\mathrm{fair}}(\theta)$ penalizes disparities expressed in terms of approval or error rates that also determine profit via $R\cdot \mathrm{TP} - C\cdot \mathrm{FP}$. At deployment time, approvals are made by thresholding $\hat{p}_\theta$ at $t^\star$. With heterogeneous (R_i, C_i), one uses the applicant-specific threshold $t_i^\star = C_i/(R_i{+}C_i)$. In summary, minimizing

$$\widehat{\mathcal{L}}(\theta;\lambda) \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big(-\,y_i \log \hat{p}_\theta(x_i,a_i) \;-\; (1-y_i)\log\big(1-\hat{p}_\theta(x_i,a_i)\big)\Big) \;+\; \lambda\, J_{\mathrm{fair}}(\theta)$$

is equivalent to learning the profit-relevant probabilities (via BCE) while trading off expected profit against a fairness cost that is explicitly tied to false-positive and false-negative behavior across groups.

D Additional algorithmic details

In the following, we provide additional details on the training algorithm of our classification model.

D.1 Preprocessing and parameterization

Let $x \in \mathbb{R}^p$ denote the input. For continuous features $i \in \mathcal{N}$ we standardize $\tilde{x}_i = (x_i - \mu_i)/\sigma_i$ using training-set moments (μ_i, σ_i), for categorical features $i \in C$ we keep the ordinal-encoded values but learn class-specific biases internally. Standardization equalizes the scale of inputs so that a common optimizer and learning rate work across features. It also stabilizes warm starts and the monotonicity and clarity penalties, which operate on comparable numeric ranges (see below).

Denote the main-effect subnetworks by $f_i(\cdot;\, \theta_i)$ and the pairwise interaction subnetworks by $f_{ij}(\cdot, \cdot;\, \theta_{ij})$ for $(i, j) \in \mathcal{I}$. The model score can be written as

$$s_\theta(x) \;=\; b \;+\; \sum_{i=1}^{p} w_i \, f_i(\tilde{x}_i;\, \theta_i) \;+\; \sum_{(i,j)\in\mathcal{I}} v_{ij} \, f_{ij}(\tilde{x}_i, \tilde{x}_j;\, \theta_{ij}),$$

with learnable scalar gates w_i and v_{ij} that control the scale (and sparsity) of each shape function. In Equation 1, we omitted this subtlety and absorbed the gate parameters into the functions f for brevity. Gates w_i and v_{ij} give the optimizer a simple way to up- or down-weight whole subnetworks. When paired with ℓ_1 penalties (see below), they naturally drive unneeded shapes to zero.

We also maintain zero-meaned shapes by centering each f_i and f_{ij} around their training set averages and compensating in the intercept b, which preserves predictions while improving identifiability. Centering each learned shape removes arbitrary offsets and prevents leakage of average effects into the intercept, making feature importances and pruning decisions more consistent.

D.2 Warm starts

To stabilize training and speed up convergence, we warm-initialize the individual subnetworks using weights from fast spline surrogates. For main effects we fit a linear GAM of the form

$$g(x) \;=\; \alpha \;+\; \sum_{i=1}^{p} g_i(\tilde{x}_i),$$

with univariate B-splines g_i (low-order, adaptively chosen number of knots). For interactions, after freezing the main effects we compute residuals (classification case): $r = \mathbb{1}\{y = 1\} - \sigma(s_\theta^{\text{main}}(x))$ and fit a GAM with tensor-product splines

$$h(x) \;=\; \beta \;+\; \sum_{(i,j)\in\mathcal{I}} h_{ij}(\tilde{x}_i, \tilde{x}_j).$$

Each neural subnetwork $\{f_i, f_{ij}\}$ is then individually pre-fit to samples from the corresponding teacher $\{g_i, h_{ij}\}$ by minimizing per-sample squared error. This defines informed initial weights (θ_i, θ_{ij}) and gates. Warm starts reduce the search space the neural nets must explore and give sensible initial shapes even with limited data. The teacher models are cheap and convex (or nearly so) in their parameters, so they capture smooth trends and simple interactions quickly. Pre-fitting each subnetwork to its teacher reduces variance at the start of optimization and leads to faster and more stable convergence, especially when combined with sparsity and monotonicity constraints.

D.2.1 Teachers for main and interaction effects

Each univariate teacher g_i is a piecewise-linear B-spline on the standardized input $z_i = \tilde{x}_i$:

$$g_i(z_i) \;=\; \sum_{m=1}^{K_i} \beta_{im}\, B_{im}(z_i), \qquad q = 1 \text{ (spline order)}.$$

We estimate $\beta_{i\cdot}$ by penalized least squares on a subsample of fixed size with weights w:

$$\min_{\beta_{i\cdot},\,\alpha} \;\sum_{n} w_n \Big(y_n^\star - \alpha - \sum_{i=1}^{p} g_i(z_{ni})\Big)^2 \;+\; \lambda \sum_{i=1}^{p} \beta_{i\cdot}^\top M_i \beta_{i\cdot},$$

where M_i is a first-difference smoothness matrix and λ is a fixed smoothness strength (we use $\lambda = 0.6$ for speed and stability). The number of basis functions per feature is reduced

as p grows,

$$K_i \;=\; \max\Big(11 - \Big\lceil \tfrac{p}{100} \Big\rceil,\; 2\Big),$$

so the teacher remains fast and low-variance when there are many predictors. Knots are placed along z_i using quantile-like positions so bins adapt to the empirical density, which improves conditioning.

For each candidate pair $(i, j) \in \mathcal{I}$ we fit a tensor-product spline to the residual target

$$r_n \;=\; \mathbb{1}\{y_n = 1\} - \sigma\big(s_\theta^{\text{main}}(x_n)\big),$$

using basis products $\{B_{im}(z_i)\, B_{jn}(z_j)\}$ and separate smoothness along each axis. The basis size scales with the number of candidate pairs,

$$K_{ij} \;=\; \max\Big(11 - \Big\lceil \tfrac{|\mathcal{I}|}{10} \Big\rceil,\; 2\Big),$$

which keeps the teacher compact when many interactions are considered. Fitting on residuals focuses the teacher on signal not already explained by the additive part.

For each feature (or pair) we draw inputs from a grid over the empirical range and evaluate the fitted teacher g_i (or h_{ij}). We then train the corresponding neural subnetwork f_i (or f_{ij}) to match these teacher values by minimizing mean squared error, using small batches and early stopping. Other inputs are held at neutral values so each subnetwork learns its own shape in isolation. The learned offsets from teachers are folded into the model intercept b, and we center the resulting shapes to have zero mean so that gates w_i and v_{ij} purely control scale.

D.3 Training objective and regularization

Let ℓ^{BCE} denote the standard binary cross-entropy. Section 2.1 already defines the equalized-odds penalty $\mathcal{P}_{\text{EO}}(\theta)$ computed on each mini-batch. This fairness term pulls class-conditional score distributions together across groups, which reduces disparities in true and false positive rates without committing to a fixed threshold.

Taking all (partly optional) additions to the algorithm into account, our total per-sample loss is

$$\begin{aligned}\ell(\theta) \;=& \ell^{\text{BCE}}(\theta) + \lambda_{\text{EO}}\, \mathcal{P}_{\text{EO}}(\theta) + \;\lambda_{\text{main}} \sum_{i=1}^{p} |w_i| \;+\; \lambda_{\text{inter}} \sum_{(i,j)\in\mathcal{I}} |v_{ij}| \\ &+\; \lambda_{\text{mono}}\, \mathcal{P}_{\text{mono}}(\theta) \;+\; \lambda_{\text{clar}}\, \mathcal{P}_{\text{clar}}(\theta)\,.\end{aligned}$$

where:

- The shape sparsity term (weighted ℓ_1 on the gates) encourages a small set of active effects and interactions, improving stability with collinear inputs and improving attribution. The gate ℓ_1 terms reduce variance and overfitting by removing weak or redundant shapes.
- The optional monotonicity penalty $\mathcal{P}_{\text{mono}}$ enforces monotone function behavior for selected features. We penalize violations of non-negativity (or non-positivity) of the directional derivatives. Writing $F(x) = s_\theta(x)$ and $\mathcal{M}^+$ (resp. $\mathcal{M}^-$) for features constrained to be increasing (resp. decreasing),

$$\mathcal{P}_{\text{mono}}(\theta) \;=\; \mathbb{E}_{x\sim\mathcal{U}(\mathcal{X})}\left[\sum_{k\in\mathcal{M}^+}\phi\left(-\frac{\partial F}{\partial x_k}(x)\right)+\sum_{k\in\mathcal{M}^-}\phi\left(\frac{\partial F}{\partial x_k}(x)\right)\right], \qquad \phi(u)=\max\{u,0\}.$$

  In practice we estimate the expectation by uniform draws from the feature data, backpropagation through $\frac{\partial F}{\partial x_k}$ using automatic differentiation, and sum the penalties. The monotonicity term can be used to encode domain knowledge that some relationships should not flip sign, which improves generalization and trust. In our main analysis we do not enforce any monotonicity constraints in order to allow for full flexibility in the estimated functional forms.
- The clarity term $\mathcal{P}_{\text{clar}}$ discourages unnecessary interaction complexity by penalizing ambiguity between main effects and interactions, making explanations cleaner by reserving interactions for genuine non-additivity. Concretely, we compute a batch-level scalar that increases when interaction subnets explain variance that is already attributable to additive structure. A simple surrogate is a covariance-style penalty between interaction outputs and main-effect residuals:

$$\widehat{\mathcal{P}}_{\text{clar}}(\theta) \;=\; \sum_{(i,j)\in\mathcal{I}} \operatorname{Var}\big[v_{ij}f_{ij}(\tilde{x}_i,\tilde{x}_j)\big] \;-\; \operatorname{Var}\big(\mathbb{E}\big[v_{ij}f_{ij}(\tilde{x}_i,\tilde{x}_j)\mid \tilde{x}_i\big]+\mathbb{E}\big[v_{ij}f_{ij}(\tilde{x}_i,\tilde{x}_j)\mid \tilde{x}_j\big]\big)\,,$$

  which is minimized when the interaction carries little purely additive signal. Our implementation uses a stable, differentiable, and more computationally simple batch surrogate, which avoids calculating the conditional expectation, with the same effect: On a minibatch $B = \{\,\tilde{x}^{(n)}\,\}_{n=1}^{N}$, the implemented clarity term is

$$\mathcal{P}_{\text{clar}}(\theta) \;=\; \sum_{(i,j)\in\mathcal{I}} \Big(\big|\widehat{\operatorname{Cov}}_B\big(f_i(\tilde{x}_i),\, v_{ij}f_{ij}(\tilde{x}_i,\tilde{x}_j)\big)\big| \;+\; \big|\widehat{\operatorname{Cov}}_B\big(f_j(\tilde{x}_j),\, v_{ij}f_{ij}(\tilde{x}_i,\tilde{x}_j)\big)\big|\Big).$$

  The absolute value induces an ℓ_1 penalty on these covariances. Minimizing $\mathcal{P}_{\text{clar}}(\theta)$ drives $v_{ij}f_{ij}(\tilde{x}_i,\tilde{x}_j)$ to be empirically orthogonal to the additive structure spanned

by $f_i(\tilde{x}_i)$ and $f_j(\tilde{x}_j)$. Consequently, the component of $\mathrm{Var}\big[v_{ij} f_{ij}(\tilde{x}_i, \tilde{x}_j)\big]$ that is explainable by $\mathbb{E}[\, v_{ij} f_{ij}(\tilde{x}_i, \tilde{x}_j) \mid \tilde{x}_i]$ and $\mathbb{E}[\, v_{ij} f_{ij}(\tilde{x}_i, \tilde{x}_j) \mid \tilde{x}_j]$ is reduced, aligning this batch surrogate with the variance-decomposition form of $\widehat{\mathcal{P}}_{\mathrm{clar}}(\theta)$ and reserving interactions for genuine non-additivity.

Mini-batch stochastic optimization uses Adam with stage-specific learning rates and early stopping.

D.4 Interaction discovery

We select pairwise interactions using a fast, binned preprocessor and an interaction detection algorithm (FAST of Lou et al. (2013)) that assigns each pair (i, j) a score S_{ij}, measuring the improvement of a pair-specific model over an additive model on the same binned representation.

Specifically, before screening interactions we discretize each feature into a small number of bins. For continuous features we use quantile-like cut points so that each bin contains a similar number of samples and for categorical features we keep the original categories (each category is one bin). Let B_i be the number of bins for feature i. The preprocessor maps every sample x_n to an integer-coded matrix

$$X \in \{0, \dots, B_1 - 1\} \times \cdots \times \{0, \dots, B_p - 1\},$$

so that $X_{n,i}$ is the bin index of sample n on feature i. This compressed, integer form supports highly efficient counting, aggregation, and table lookups, which is what the interaction detector uses internally.

On the binned scale we can define a purely additive (no interaction) model that depends only on features i and j:

$$\hat{y}_n^{\mathrm{add}} \;=\; \mu \;+\; g_i\big(X_{n,i}\big) \;+\; g_j\big(X_{n,j}\big),$$

where $g_i : \{0, \dots, B_i - 1\} \to \mathbb{R}$ and $g_j : \{0, \dots, B_j - 1\} \to \mathbb{R}$ are univariate shape tables (one score per bin), i.e., y is explained using only per-bin offsets for i and j. The interaction-augmented counterpart adds a two-way table

$$\hat{y}_n^{\mathrm{pair}} \;=\; \mu \;+\; g_i\big(X_{n,i}\big) \;+\; g_j\big(X_{n,j}\big) \;+\; h_{ij}\big(X_{n,i}, X_{n,j}\big),$$

where h_{ij} assigns an extra score to each bin pair $(b_i, b_j) \in \{0, \dots, B_i - 1\} \times \{0, \dots, B_j - 1\}$.

Given the binned data X, the labels y, optional sample weights w, and the current

main-effect scores from the univariate fitting stage, our variant of the FAST algorithm computes how much additional predictive signal is captured by allowing a two-way table for (i, j) beyond the best additive fit on i and j alone. Concretely, for each pair (i, j) the detector performs the following operations on the same binned grid:

1. Form sufficient statistics per bin (and per class for classification), e.g., weighted counts and weighted response sums in each i-bin, each j-bin, and each (i, j) bin-pair.
2. Fit the best additive tables (g_i, g_j) on $\{i, j\}$ by minimizing the chosen loss on bins (squared error for regression, cross-entropy–style terms for classification) subject to simple leaf-size constraints.
3. Fit the best additive-plus-pair model (g_i, g_j, h_{ij}) on the same binned data and constraints.
4. Compute the hold-out loss values $\mathcal{L}^{ij}_{\text{add}}$ and $\mathcal{L}^{ij}_{\text{pair}}$ on exactly the same representation.

The interaction score is the improvement:

$$S_{ij} \;=\; \mathcal{L}^{ij}_{\text{add}} \;-\; \mathcal{L}^{ij}_{\text{pair}} \;\geq 0,$$

so larger S_{ij} means the pair table h_{ij} explains structure that cannot be replicated by separate g_i and g_j. Because everything is computed on integer bins, these fits reduce to fast table updates rather than expensive neural training.

Additionally, heredity is enforced by restricting candidates to those where at least one of i or j is already active as a main effect. The top K pairs (bounded by $\binom{p}{2}$ and by a hyperparameter) become $\mathcal{I}$. Binning creates a compact representation that makes exhaustive pair screening computationally feasible. Comparing a pair model against an additive baseline isolates true interaction signal, while the heredity rule shrinks the search space and reduces spurious pairs by requiring that interactions involve features that already matter on their own. The capacity parameter K controls complexity and training cost.

D.5 Pruning

In addition to the in-processing regularization, we encourage sparsity through pruning: After training main effects (stage 1), we rank them by a variance-weighted gate magnitude

$$\text{scale}_i \;=\; |w_i|^2 \cdot \text{Var}\big[f_i(\tilde{x}_i)\big]\,,$$

computed on the training set. We then build a forward path by activating effects in descending order and record validation loss at each step: Let $\widehat{\mathcal{L}}^{\text{val}}(k)$ be the validation

loss with the top k main effects enabled. We select

$$k^\star \;=\; \min\left\{k : \frac{\widehat{\mathcal{L}}^{\text{val}}(k)}{\min_m \widehat{\mathcal{L}}^{\text{val}}(m)} - 1 \;<\; \delta\right\},$$

with tolerance $\delta \geq 0$ (loss threshold) to favor sparser models when performance is similar. Interactions are pruned analogously using $\text{scale}_{ij} = |v_{ij}|^2 \cdot \text{Var}\big[f_{ij}(\tilde{x}_i, \tilde{x}_j)\big]$ and a similar validation procedure. The scales reflect both the learned gate and the variability of each shape, so they capture contribution to prediction spread. The forward path mimics a regularization path but uses the trained network components directly, which avoids retraining at each sparsity level. The tolerance δ implements the idea that ties on validation should be broken in favor of simpler models.

D.6 Three-stage optimization and early stopping

The full training procedure is: (i) train main-effect blocks only (interaction blocks frozen), (ii) add discovered interactions and train interaction blocks only (main effects frozen), (iii) jointly fine-tune all parameters with a reduced learning rate.

They are implemented in the following way:

1. Main (univariate) effect learning: For each feature i, a neural subnetwork f_i is trained while minimizing the chosen loss function. Each subnet consists of one input neuron for x_i, a hidden layer of 20 units, and an output neuron. All subnets use ReLU activations and are trained in parallel via gradient descent. Training stops after a pre-defined number of epochs or when performance stops improving (as defined by an early stopping threshold).
2. Interaction learning: After selecting the interaction pairs using the FAST algorithm (see Lou et al., 2013), neural subnetworks f_{ij} are trained while keeping the main effect subnetworks fixed. The interaction subnetworks have two input neurons (for x_i and x_j), two hidden layers of 20 neurons each, and one output neuron. Otherwise, training proceeds as in the first stage.
3. Joint fine-tuning: Once all main effect and interaction subnetworks are trained, a global optimization step is performed in which all parameters are updated simultaneously (with a lower learning rate in order to avoid losing information from the previous training steps). Due to the structural separability of the neural subnetworks, this stage does not pollute the interpretability of individual shape functions, but allows for small coordinated adjustments that improve performance.

Each stage performs at most a fixed number of epochs with an auto early-stopping patience

$$\text{patience} \;=\; \max\Big(5,\; \min\big(\lfloor 5000\, p/(I_{\text{per-epoch}} \cdot B)\rfloor,\; 100\big)\Big),$$

where p is the number of features, B the batch size, and $I_{\text{per-epoch}}$ the capped number of mini-batches per epoch. During the validation process we also allow different fixed numbers of rounds as options for the patience hyperparameter.

When monotonicity is active, we do not impose hard constraints, but instead apply a soft, differentiable penalty during training and periodically check whether violations remain. Concretely, whenever early stopping would halt a stage, we run a simple check to see if the model is (approximately) monotone in the features that were marked as increasing or decreasing. If violations are still present, we increase the monotonicity weight λ_{mono} and continue training. Otherwise the model is accepted.

D.7 Inference details

At inference, continuous features are normalized by (μ_i, σ_i) and values are clipped to the training range when enabled. These steps avoid uncontrolled extrapolation of shape functions outside the observed range and keep the same normalization as in training to preserve calibration. Predictions use the full model s_θ with the selected effect sets, probabilities are $\hat{p}(x) = \sigma(s_\theta(x))$. The centering of effects guarantees that changes in w_i or v_{ij} scale shapes without inadvertently shifting the overall intercept. In combination with the gates, centering ensures that turning an effect on or off changes only the intended component, which makes ablations and explanations faithful.

The model outputs probabilities $\hat{p}(x) = \sigma(s_\theta(x))$, while downstream decisions often require a hard rule. In practice, we either fix a threshold $t \in (0, 1)$ and predict $\hat{y} = \mathbb{1}\{\hat{p}(x) \geq t\}$ (for example, our main analysis uses a threshold of $t = 0.5$), or we set t to meet a fixed rate, e.g., accept top q% (see, for example, the lender capacity matching in Section 5.3). Because the fairness term aligns class-conditional score distributions across groups in a threshold-free way, improvements usually transfer across a range of thresholds (see above).

E Omitted proofs

In the following, we present proofs supporting the chosen penalty term. Specifically, we show that (1) the soft penalty term is equivalent to a weighted average over thresholds, (2) the penalty is bounded by a weighted EO disparity, and (3) the penalty generalizes out-of-sample. We summarize the notation in the following way: Let $X \in \mathcal{X}$ denote features, $Y \in \{0,1\}$ the outcome, and $A \in \mathcal{G}$ a protected attribute. Let $S = S_\theta(X) = s_\theta(X) \in \mathbb{R}$ be the model score, and $h_\theta(x) = \sigma\big(s_\theta(x)\big)$ the predicted probability with $\sigma(u) = (1 + e^{-u})^{-1}$.

For each group-label cell $(g, y) \in \mathcal{G} \times \{0,1\}$, define

$$\mu_{g,y}(\theta) \;:=\; \mathbb{E}\big[h_\theta(X)\,\big|\,A{=}g,\, Y{=}y\big] \;=\; \mathbb{E}\big[\sigma(S_\theta)\,\big|\,A{=}g,\, Y{=}y\big] \quad \in [0,1].$$

For a collection $(v_g)_{g\in\mathcal{G}}$, we write

$$\mathrm{Var}_{g\in\mathcal{G}}(v_g) \;:=\; \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\big(v_g - \bar{v}\big)^2, \qquad \bar{v} := \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} v_g,$$

i.e., the uniform variance across groups.[27]

The population fairness penalty implemented in training is the sum of cross-group variances of these conditional means:

$$\mathcal{P}_{\mathrm{mean}}(\theta) \;:=\; \mathrm{Var}_{g\in\mathcal{G}}\big(\mu_{g,1}(\theta)\big) \;+\; \mathrm{Var}_{g\in\mathcal{G}}\big(\mu_{g,0}(\theta)\big).$$

On a sample $\{(X_\ell, Y_\ell, A_\ell)\}_{\ell=1}^n$, write $n_{g,y} := \sum_{\ell=1}^n \mathbf{1}\{A_\ell{=}g,\, Y_\ell{=}y\}$ and, for cells with $n_{g,y} \geq 1$,

$$\widehat{\mu}_{g,y}(\theta) \;:=\; \frac{1}{n_{g,y}} \sum_{\ell:\, A_\ell=g,\, Y_\ell=y} h_\theta(X_\ell).$$

Accordingly, the empirical penalty is

$$\widehat{\mathcal{P}}_{\mathrm{mean}}(\theta) \;=\; \mathrm{Var}_{g\in\mathcal{G}}\big(\widehat{\mu}_{g,1}(\theta)\big) \;+\; \mathrm{Var}_{g\in\mathcal{G}}\big(\widehat{\mu}_{g,0}(\theta)\big).$$

(We assume throughout that every group-label cell used in the empirical penalty has $n_{g,y} \geq 1$.)

Further, let $R^\theta_{g,y}(\tau) := \Pr\big(S_\theta \geq \tau \mid A{=}g,\, Y{=}y\big)$ be the rate/survival curve of S_θ inside cell (g, y) dependent on threshold τ, and let $w(\tau) := \sigma'(\tau) = \sigma(\tau)\big(1 - \sigma(\tau)\big)$. Note that

[27] A weighted alternative using group (or cell) probabilities can be used instead. Our choice here matches the empirical implementation above.

$w(\tau) \geq 0$ for all τ and

$$\int_{-\infty}^{\infty} w(\tau)\, d\tau \;=\; \int_{-\infty}^{\infty} \sigma'(\tau)\, d\tau \;=\; \lim_{b\to\infty} \sigma(b) - \lim_{a\to-\infty} \sigma(a) \;=\; 1,$$

so w is a probability density on $\mathbb{R}$.

Theorem 1 (Tonelli's theorem, nonnegative case)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let λ denote Lebesgue measure on $\mathbb{R}$, and let $f : \mathbb{R} \times \Omega \to [0, \infty]$ be measurable. Then*

$$\int_{\Omega}\left(\int_{\mathbb{R}} f(\tau,\omega)\, d\lambda(\tau)\right) d\mathbb{P}(\omega) \;=\; \int_{\mathbb{R}}\left(\int_{\Omega} f(\tau,\omega)\, d\mathbb{P}(\omega)\right) d\lambda(\tau) \;=\; \int_{\mathbb{R}\times\Omega} f\, d(\lambda\otimes\mathbb{P}).$$

In particular, the order of integration may be interchanged without any integrability assumptions beyond non-negativity.

Application in Proposition 1. We apply Theorem 1 with

$$f(\tau,\omega) \;=\; \sigma'(\tau)\,\mathbf{1}\{\tau \leq S_\theta(\omega)\} \;\geq 0,$$

so that

$$\mathbb{E}\left[\int_{-\infty}^{\infty} \sigma'(\tau)\,\mathbf{1}\{\tau \leq S_\theta\}\, d\tau\right] \;=\; \int_{-\infty}^{\infty} \sigma'(\tau)\,\mathbb{E}[\mathbf{1}\{\tau \leq S_\theta\}]\, d\tau.$$

E.1 Proof of Proposition 1

Proof. Fix (g, y) and, for brevity, write all expectations and probabilities *conditionally on* $(A{=}g, Y{=}y)$. We aim to show

$$\mathbb{E}\big[\sigma(S_\theta)\big] \;=\; \int_{-\infty}^{\infty} \sigma'(\tau)\, \Pr(S_\theta \geq \tau)\, d\tau \;=\; \int w(\tau)\, R^{\theta}_{g,y}(\tau)\, d\tau.$$

Step 1. Since σ is absolutely continuous with derivative σ', and $\lim_{t\to-\infty} \sigma(t) = 0$, we have for each $s \in \mathbb{R}$,

$$\sigma(s) \;=\; \int_{-\infty}^{s} \sigma'(\tau)\, d\tau \;=\; \int_{-\infty}^{\infty} \sigma'(\tau)\,\mathbf{1}\{\tau \leq s\}\, d\tau.$$

Applying this identity to $s = S_\theta$ and taking expectations yields

$$\mathbb{E}[\sigma(S_\theta)] \;=\; \mathbb{E}\left[\int_{-\infty}^{\infty} \sigma'(\tau)\,\mathbf{1}\{\tau \leq S_\theta\}\, d\tau\right].$$

Step 2. The integrand $\sigma'(\tau)\mathbf{1}\{\tau \le S_\theta\}$ is nonnegative for each (τ, S_θ), hence Tonelli's theorem (Theorem 1, see, e.g., Folland (1999) for more details) applies directly:

$$\mathbb{E}[\sigma(S_\theta)] \;=\; \int_{-\infty}^{\infty} \sigma'(\tau)\,\mathbb{E}\big[\mathbf{1}\{\tau \le S_\theta\}\big]\,d\tau.$$

Step 3. For each fixed τ, $\mathbb{E}[\mathbf{1}\{\tau \le S_\theta\}] = \Pr(S_\theta \ge \tau) = R^\theta_{g,y}(\tau)$. Therefore,

$$\mathbb{E}[\sigma(S_\theta)] \;=\; \int_{-\infty}^{\infty} \sigma'(\tau)\,\Pr(S_\theta \ge \tau)\,d\tau \;=\; \int w(\tau)\,R^\theta_{g,y}(\tau)\,d\tau.$$

This completes the proof. □

Corollary 1 (Penalty equals variance of w-averaged rates)**.** *For $y \in \{0,1\}$, define the w-weighted expectation $\mathbb{E}_w[\cdot] := \int_{-\infty}^{\infty} w(\tau)\,(\cdot)\,d\tau$. Then, for each (g, y),*

$$\mu_{g,y}(\theta) \;=\; \mathbb{E}\big[\sigma(S_\theta) \mid A{=}g, Y{=}y\big] \;=\; \int w(\tau)\,R^\theta_{g,y}(\tau)\,d\tau \;=\; \mathbb{E}_w\big[R^\theta_{g,y}\big],$$

and hence

$$\mathcal{P}_{\text{mean}}(\theta) \;=\; \mathrm{Var}_{g\in\mathcal{G}}\big(\mathbb{E}_w[R^\theta_{g,1}]\big) \;+\; \mathrm{Var}_{g\in\mathcal{G}}\big(\mathbb{E}_w[R^\theta_{g,0}]\big).$$

E.2 Proof of Proposition 2

Proof. Fix $y \in \{0,1\}$ and write

$$m_g \;:=\; \mathbb{E}_w[R^\theta_{g,y}], \qquad \bar{R}^\theta_y(\tau) \;:=\; \frac{1}{G}\sum_{g=1}^{G} R^\theta_{g,y}(\tau), \quad G = |\mathcal{G}|.$$

Recall our convention that Var_g denotes the uniform variance across $g \in \{1,\dots,G\}$. Then

$$\mathrm{Var}_g(m_g) \;=\; \frac{1}{G}\sum_{g=1}^{G}\Big(m_g - \frac{1}{G}\sum_{h=1}^{G} m_h\Big)^2 \;=\; \frac{1}{G}\sum_{g=1}^{G}\Big(\int w(\tau)\,[R^\theta_{g,y}(\tau) - \bar{R}^\theta_y(\tau)]\,d\tau\Big)^2.$$

Let $\Delta_g(\tau) := R^\theta_{g,y}(\tau) - \bar{R}^\theta_y(\tau)$. Since $w(\tau) \ge 0$ and $\int w = 1$, $w(\tau)\,d\tau$ is a probability measure. By Jensen's inequality for the convex map $z \mapsto z^2$,

$$\Big(\int w(\tau)\,\Delta_g(\tau)\,d\tau\Big)^2 \;\le\; \int w(\tau)\,\Delta_g(\tau)^2\,d\tau.$$

Averaging over g and interchanging the (finite) sum and the integral (or, equivalently, applying Tonelli's theorem (Theorem 1) to the nonnegative integrand $\Delta_g(\tau)^2$) gives

$$\mathrm{Var}_g(m_g) \;\le\; \int w(\tau)\,\frac{1}{G}\sum_{g=1}^{G}\Delta_g(\tau)^2\,d\tau \;=\; \int w(\tau)\;\mathrm{Var}_g\big(R^{\theta}_{g,y}(\tau)\big)\,d\tau.$$

Applying this bound separately for $y = 1$ and $y = 0$ and recalling from Corollary 1 that

$$\mathcal{P}_{\mathrm{mean}}(\theta) \;=\; \mathrm{Var}_g\big(\mathbb{E}_w[R^{\theta}_{g,1}]\big) \;+\; \mathrm{Var}_g\big(\mathbb{E}_w[R^{\theta}_{g,0}]\big),$$

we obtain

$$\mathrm{Var}_g\big(\mathbb{E}_w[R^{\theta}_{g,1}]\big) \;\le\; \int w(\tau)\;\mathrm{Var}_g\big(R^{\theta}_{g,1}(\tau)\big)\,d\tau, \qquad \mathrm{Var}_g\big(\mathbb{E}_w[R^{\theta}_{g,0}]\big) \;\le\; \int w(\tau)\;\mathrm{Var}_g\big(R^{\theta}_{g,0}(\tau)\big)\,d\tau.$$

Adding these inequalities and using linearity of the integral yields

$$\mathcal{P}_{\mathrm{mean}}(\theta) \;\le\; \underbrace{\int w(\tau)\;\mathrm{Var}_g\big(R^{\theta}_{g,1}(\tau)\big)\,d\tau}_{\mathcal{D}_1(\theta)} \;+\; \underbrace{\int w(\tau)\;\mathrm{Var}_g\big(R^{\theta}_{g,0}(\tau)\big)\,d\tau}_{\mathcal{D}_0(\theta)}.$$

Therefore:

$$\mathcal{P}_{\mathrm{mean}}(\theta) = \mathrm{Var}_g\big(\mathbb{E}_w[R^{\theta}_{g,1}]\big) + \mathrm{Var}_g\big(\mathbb{E}_w[R^{\theta}_{g,0}]\big) \;\le\; \mathcal{D}_1(\theta) + \mathcal{D}_0(\theta)\,.$$

□

E.3 Proof of Proposition 3

We show that $\widehat{\mathcal{P}}_{\mathrm{mean}}(\theta)$ uniformly concentrates around $\mathcal{P}_{\mathrm{mean}}(\theta)$ as $n \to \infty$ under standard capacity control for the additive class.

Function classes. Let $\mathcal{H} := \{h_\theta = \sigma \circ s_\theta : \theta \in \Theta\}$ be the probability predictors, bounded in $[0, 1]$. Assume the score class admits the additive, L^1-gated representation

$$\mathcal{S} \;=\; \left\{ x \mapsto \beta_0 + \sum_i \alpha_i f_i(x_i) + \sum_{(i,j)\in\mathcal{I}} \alpha_{ij} f_{ij}(x_i, x_j) : \sum_i |\alpha_i| \le \Lambda_1, \sum_{(i,j)} |\alpha_{ij}| \le \Lambda_2 \right\},$$

with base functions satisfying $\|f_i\|_\infty \le B_1$ (values never exceed B_1), $\mathrm{Lip}(f_i) \le L_1$ (slope $\le L_1$), $\|f_{ij}\|_\infty \le B_2$, $\mathrm{Lip}(f_{ij}) \le L_2$. Let M bound the number of active components (main effects plus interactions).

For any measurable $h : \mathcal{X} \to [0,1]$, write $P_{g,y}h := \mathbb{E}[h(X) \mid A{=}g, Y{=}y]$ and $\widehat{P}_{g,y}h := \frac{1}{n_{g,y}} \sum_{\ell \in I_{g,y}} h(X_\ell)$. Therefore, $(\widehat{P}_{g,y} - P_{g,y})h$ denotes the empirical–population mean gap for h in cell (g,y).

Cell masses. Assume minimal cell mass:

$$\Pr(A{=}g,\ Y{=}y) \ge p_{\min} > 0 \quad \text{for all } (g,y) \in \mathcal{G} \times \{0,1\}. \tag{11}$$

Theorem 2 (McDiarmid (1989)'s inequality (bounded differences)). *Let $Z_1, \dots, Z_m$ be independent random variables taking values in sets $\mathcal{Z}_1, \dots, \mathcal{Z}_m$. Let $F : \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_m \to \mathbb{R}$ satisfy the bounded-differences property: for some constants $c_1, \dots, c_m \ge 0$,*

$$\sup_{z_1,\dots,z_m,z_i'} \left|F(z_1,\dots,z_i,\dots,z_m) - F(z_1,\dots,z_i',\dots,z_m)\right| \ \le\ c_i \quad \textit{for each } i.$$

Then, for all $t > 0$,

$$\Pr(F - \mathbb{E}[F] \ge t) \ \le\ \exp\left(-\frac{2t^2}{\sum_{i=1}^m c_i^2}\right), \qquad \Pr(|F - \mathbb{E}[F]| \ge t) \ \le\ 2\exp\left(-\frac{2t^2}{\sum_{i=1}^m c_i^2}\right).$$

Lemma 1. *There exist absolute constants $c_1, c_2 > 0$ such that, with probability at least $1-\delta$,*

$$\max_{(g,y)} \sup_{\theta \in \Theta} \left|\widehat{\mu}_{g,y}(\theta) - \mu_{g,y}(\theta)\right| \ \le\ c_1 \max_{(g,y)} \widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H}) \ +\ c_2 \max_{(g,y)} \sqrt{\frac{\log(2G/\delta)}{n_{g,y}}},$$

where $\widehat{\mathfrak{R}}_m(\mathcal{H})$ is the empirical Rademacher complexity of $\mathcal{H}$ on m points.

Proof. Fix (g,y) and condition on the index set $I_{g,y} := \{\ell : A_\ell{=}g,\ Y_\ell{=}y\}$ of size $n_{g,y}$; then $\{X_\ell : \ell \in I_{g,y}\}$ are i.i.d. from $(X \mid A{=}g, Y{=}y)$. Write $P_{g,y}$ for the conditional distribution and $\widehat{P}_{g,y}$ the empirical measure on $\{X_\ell : \ell \in I_{g,y}\}$. We bound the uniform deviation $\sup_{h \in \mathcal{H}} |(\widehat{P}_{g,y} - P_{g,y})h|$.

(i) Symmetrization. Let $\{\varepsilon_\ell\}_{\ell \in I_{g,y}}$ be i.i.d. Rademacher random variables independent of

the data.[28] Then

$$\mathbb{E}\Big[\sup_{h\in\mathcal{H}}|(\widehat{P}_{g,y}-P_{g,y})h|\;\Big|\;I_{g,y}\Big]\;\le\;2\,\mathbb{E}\Big[\sup_{h\in\mathcal{H}}\Big|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell h(X_\ell)\Big|\;\Big|\;I_{g,y}\Big]\;=\;2\,\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H}).$$

The proof of the above inequality is as follows. Let

$$P_{g,y,h}\;:=\;\mathbb{E}\big[h(X)\mid A=g,\,Y=y\big]\,,\qquad \widehat{P}_{g,y,h}\;:=\;\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}h(X_\ell),$$

and let $(X'_1,\dots,X'_{n_{g,y}})$ be an independent *ghost sample* (i.e., a fresh, independent copy of the data) from the same cell distribution $P_{g,y}$. Then

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|(\widehat{P}_{g,y}-P_{g,y})h\right|\,\middle|\,I_{g,y}\right]\le\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\big(h(X'_\ell)-h(X_\ell)\big)\right|\,\middle|\,I_{g,y}\right].$$

(We replaced the unknown population mean by the ghost empirical mean; Jensen gives the "≤".)

Now introduce Rademacher signs $\{\varepsilon_\ell\}_{\ell\in I_{g,y}}$, i.i.d. with $\varepsilon_\ell\in\{\pm1\}$ and independent of the data. Conditioned on the paired sample $\{(X_\ell,X'_\ell)\}$, the vector $\big(h(X'_\ell)-h(X_\ell)\big)_{\ell\in I_{g,y}}$ is deterministic. Using the symmetry of ε and the fact that (X_ℓ,X'_ℓ) and (X'_ℓ,X_ℓ) have the same joint law, one obtains the standard symmetrization step:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\big(h(X'_\ell)-h(X_\ell)\big)\right|\,\middle|\,I_{g,y}\right]\le\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell\big(h(X'_\ell)-h(X_\ell)\big)\right|\,\middle|\,I_{g,y}\right].$$

Apply the triangle inequality inside the supremum:

$$\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell\big(h(X'_\ell)-h(X_\ell)\big)\right|\;\le\;\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell h(X'_\ell)\right|+\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell h(X_\ell)\right|.$$

Taking the supremum over h and expectations (still conditional on $I_{g,y}$),

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\big(h(X'_\ell)-h(X_\ell)\big)\right|\,\middle|\,I_{g,y}\right]\le\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell h(X'_\ell)\right|+\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell h(X_\ell)\right|\,\middle|\,I_{g,y}\right].$$

[28] Note the Rademacher signs $\varepsilon_i\in\{+1,-1\}$ are an auxiliary device we introduce to "randomize the sign" in a way that centers the process for the symmetrization step. They're independent of everything (data and labels).

Since X'_ℓ and X_ℓ have the same distribution,

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\big(h(X'_\ell)-h(X_\ell)\big)\right|\,\middle|\,I_{g,y}\right]\le 2\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell h(X_\ell)\right|\,\middle|\,I_{g,y}\right]=2\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H}).$$

Combining the displays yields exactly the desired inequality:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|(\widehat{P}_{g,y}-P_{g,y})h\right|\,\middle|\,I_{g,y}\right]\;\le\;2\,\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}\varepsilon_\ell h(X_\ell)\right|\,\middle|\,I_{g,y}\right]=2\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H}).$$

(ii) Concentration via McDiarmid. Define, for the fixed cell (g,y) and its index set $I_{g,y}$ of size $n_{g,y}$,

$$F\big((X_\ell)_{\ell\in I_{g,y}}\big)\;:=\;\sup_{h\in\mathcal{H}}\left|(\widehat{P}_{g,y}-P_{g,y})h\right|\;=\;\sup_{h\in\mathcal{H}}\left|\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}h(X_\ell)\;-\;\mathbb{E}[h(X)\mid A{=}g,Y{=}y]\right|.$$

Since $h:\mathcal{X}\to[0,1]$, changing a single X_ℓ can change the average $\frac{1}{n_{g,y}}\sum_{\ell\in I_{g,y}}h(X_\ell)$ by *at most* $1/n_{g,y}$ for every h, and hence it can change the supremum over h by at most $1/n_{g,y}$ as well. Therefore F has bounded differences with constants $c_\ell=1/n_{g,y}$ for each $\ell\in I_{g,y}$, so that $\sum_{\ell\in I_{g,y}}c_\ell^2=n_{g,y}\cdot(1/n_{g,y})^2=1/n_{g,y}$. Conditional on $I_{g,y}$, the variables $\{X_\ell:\ell\in I_{g,y}\}$ are independent draws from $P_{g,y}(\cdot)=\Pr(X\in\cdot\mid A{=}g,Y{=}y)$, so McDiarmid's inequality (Theorem 2) applies and yields, for all $t>0$,

$$\Pr\Big(F-\mathbb{E}[F\mid I_{g,y}]\ge t\,\Big|\,I_{g,y}\Big)\;\le\;\exp\big(-2\,n_{g,y}\,t^2\big),\;\Pr\Big(|F-\mathbb{E}[F\mid I_{g,y}]|\ge t\,\Big|\,I_{g,y}\Big)\;\le\;2\,\exp\big(-2\,n_{g,y}\,t^2\big).$$

Combining this with the symmetrization bound from (i), $\mathbb{E}[F\mid I_{g,y}]\le 2\,\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H})$, gives that with probability at least $1-\delta_g$ (conditional on $I_{g,y}$),

$$\sup_{h\in\mathcal{H}}\left|(\widehat{P}_{g,y}-P_{g,y})h\right|\;=\;F\;\le\;2\,\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H})\;+\;\sqrt{\frac{\log(2/\delta_g)}{2\,n_{g,y}}}.$$

(iii) Union bound across g and y.

Choose equal budgets $\delta_{g,y}=\delta/(2G)$ for each g at each fixed label y. A union bound over $g=1,\dots,G$ (holding y fixed) yields, with probability at least $1-\delta/2$

$$\max_g\;\sup_{\theta\in\Theta}\left|\widehat{\mu}_{g,y}(\theta)-\mu_{g,y}(\theta)\right|\;\le\;2\,\max_g\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H})+\max_g\sqrt{\tfrac{\log(4G/\delta)}{2n_{g,y}}}.$$

A union bound over $y \in \{0, 1\}$ (two labels) gives with a probability of at least $1 - \delta$

$$\max_{(g,y)} \sup_{\theta \in \Theta} \left| \widehat{\mu}_{g,y}(\theta) - \mu_{g,y}(\theta) \right| \;\le\; 2 \max_{(g,y)} \widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H}) + \max_{(g,y)} \sqrt{\frac{\log(4G/\delta)}{2n_{g,y}}}.$$

Absorbing the constants into absolute constants $c_1, c_2 > 0$ gives the result in Lemma 1. Finally, since $\widehat{\mu}_{g,y}(\theta) = \widehat{P}_{g,y} h_\theta$ and $\mu_{g,y}(\theta) = P_{g,y} h_\theta$, this represents a bound for the empirical-population mean gap for the cell (g, y). □

Lemma 2. *The logistic is* $1/4$*-Lipschitz, therefore*

$$\widehat{\mathfrak{R}}_m(\mathcal{H}) \;\le\; \frac{1}{4} \widehat{\mathfrak{R}}_m(\mathcal{S}).$$

Proof. For any sample $\{x_\ell\}_{\ell=1}^m$ and Rademacher random variables $\{\varepsilon_\ell\}$,

$$\widehat{\mathfrak{R}}_m(\mathcal{H}) = \mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \frac{1}{m} \sum_{\ell=1}^m \varepsilon_\ell \, \sigma\big(s_\theta(x_\ell)\big) \right], \quad \widehat{\mathfrak{R}}_m(\mathcal{S}) = \mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \frac{1}{m} \sum_{\ell=1}^m \varepsilon_\ell \, s_\theta(x_\ell) \right],$$

Let $\tilde{\sigma}(u) := \sigma(u) - \sigma(0)$ so that $\tilde{\sigma}(0) = 0$ and $\tilde{\sigma}$ is also $1/4$–Lipschitz.[29] Then, for each realization of ε,

$$\begin{aligned} \sup_\theta \frac{1}{m} \sum_\ell \varepsilon_\ell \, \sigma(s_\theta(x_\ell)) &= \sup_\theta \left[\frac{1}{m} \sum_\ell \varepsilon_\ell \, \tilde{\sigma}(s_\theta(x_\ell)) \;+\; \sigma(0) \frac{1}{m} \sum_\ell \varepsilon_\ell \right] \\ &= \sup_\theta \frac{1}{m} \sum_\ell \varepsilon_\ell \, \tilde{\sigma}(s_\theta(x_\ell)) \;+\; \sigma(0) \frac{1}{m} \sum_\ell \varepsilon_\ell, \end{aligned}$$

because the last term does not depend on θ (so it factors out of the supremum). Taking expectations in ε and using $\mathbb{E}_\varepsilon[\sum_\ell \varepsilon_\ell] = 0$,

$$\widehat{\mathfrak{R}}_m(\mathcal{H}) = \mathbb{E}_\varepsilon \left[\sup_\theta \frac{1}{m} \sum_\ell \varepsilon_\ell \, \tilde{\sigma}(s_\theta(x_\ell)) \right].$$

Talagrand's (vector) contraction lemma states: if $\phi : \mathbb{R} \to \mathbb{R}$ is L–Lipschitz with $\phi(0) = 0$, then

$$\mathbb{E}_\varepsilon \Big[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_\ell \varepsilon_\ell \, \phi\big(f(x_\ell)\big) \Big] \;\le\; L \, \mathbb{E}_\varepsilon \Big[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_\ell \varepsilon_\ell \, f(x_\ell) \Big].$$

[29] Note $1/4$-Lipschitz of $\sigma(x)$ comes from the fact that $\sigma'(x) = \sigma(x)(1 - \sigma(x)) \le \frac{1}{4}$, so $|\sigma(x) - \sigma(y)| \le \frac{1}{4}|x - y|$. And, since $\tilde{\sigma}(x)$ is a constant shift of $\sigma(x)$, that is also $1/4$-Lipschitz.

Apply this with $\phi = \tilde{\sigma}$ (so $L = 1/4$) and $\mathcal{F} = \mathcal{S}$ to obtain

$$\mathbb{E}_\varepsilon\left[\sup_\theta \frac{1}{m}\sum_\ell \varepsilon_\ell\, \tilde{\sigma}\big(s_\theta(x_\ell)\big)\right] \;\le\; \frac{1}{4}\,\mathbb{E}_\varepsilon\left[\sup_\theta \frac{1}{m}\sum_\ell \varepsilon_\ell\, s_\theta(x_\ell)\right] = \frac{1}{4}\,\widehat{\mathfrak{R}}_m(\mathcal{S}).$$

Combining the previous results gives

$$\widehat{\mathfrak{R}}_m(\mathcal{H}) = \mathbb{E}_\varepsilon\left[\sup_\theta \frac{1}{m}\sum_\ell \varepsilon_\ell\, \tilde{\sigma}(s_\theta(x_\ell))\right] \;\le\; \frac{1}{4}\,\widehat{\mathfrak{R}}_m(\mathcal{S}),$$

which is the desired inequality.

□

Lemma 3. *There exist constants $C_1, C_2 > 0$ such that*

$$\widehat{\mathfrak{R}}_m(\mathcal{S}) \;\le\; C_1\,(\Lambda_1 B_1 + \Lambda_2 B_2)\sqrt{\frac{\log M}{m}} \;+\; C_2\,(\Lambda_1 L_1 + \Lambda_2 L_2)\,\frac{1}{\sqrt{m}}.$$

Proof. Write $s(x) = \sum_{j=1}^M \alpha_j \phi_j(x)$ where the active dictionary $\{\phi_j\}$ consists of the chosen f_i's and f_{ij}'s, with $\|\alpha\|_1 \le \Lambda_1$ on the main-effect block and $\le \Lambda_2$ on the interaction block, and $\|\phi_j\|_\infty \le B$ (blockwise B_1, B_2).

Condition on the sample $\{x_\ell\}_{\ell=1}^m$ and write the active dictionary as $\phi(x) = (\phi_1(x), \dots, \phi_M(x))$. For one block with $\|\alpha\|_1 \le \Lambda$ and $|\phi_j(x)| \le B$, we have

$$\begin{aligned}
\widehat{\mathfrak{R}}_m(\mathcal{S}) &= \mathbb{E}_\varepsilon\left[\sup_{\|\alpha\|_1\le\Lambda} \frac{1}{m}\sum_{\ell=1}^m \varepsilon_\ell \sum_{j=1}^M \alpha_j \phi_j(x_\ell)\right] \\
&= \mathbb{E}_\varepsilon\left[\sup_{\|\alpha\|_1\le\Lambda} \sum_{j=1}^M \alpha_j \underbrace{\left(\frac{1}{m}\sum_{\ell=1}^m \varepsilon_\ell\, \phi_j(x_\ell)\right)}_{=:\, v_j(\varepsilon)}\right] \;=\; \mathbb{E}_\varepsilon\left[\sup_{\|\alpha\|_1\le\Lambda} \langle \alpha, v(\varepsilon)\rangle\right],
\end{aligned}$$

where $v(\varepsilon) = (v_1(\varepsilon), \dots, v_M(\varepsilon)) \in \mathbb{R}^M$. By ℓ_1–ℓ_∞ duality,

$$\sup_{\|\alpha\|_1\le\Lambda} \langle \alpha, v\rangle \;=\; \Lambda\,\|v\|_\infty, \quad \text{and in particular} \quad \sup_{\|\alpha\|_1\le\Lambda} \Big|\langle \alpha, v\rangle\Big| \;=\; \Lambda\,\|v\|_\infty,$$

(the equality is attained by choosing $\alpha = \Lambda\, e_{j^\star}\,\mathrm{sign}(v_{j^\star})$ with $j^\star \in \arg\max_j |v_j|$, where $e_{j^\star}$ is the standard basis vector with a 1 in coordinate $j^\star$ and zeros elsewhere). Therefore, we

can write

$$\widehat{\mathfrak{R}}_m(\mathcal{S}) \;=\; \Lambda\,\mathbb{E}_\varepsilon\left[\left\|\frac{1}{m}\sum_{\ell=1}^{m}\varepsilon_\ell\,\phi(x_\ell)\right\|_\infty\right] \;=\; \Lambda\,\mathbb{E}_\varepsilon\left[\max_{1\le j\le M}\left|\frac{1}{m}\sum_{\ell=1}^{m}\varepsilon_\ell\,\phi_j(x_\ell)\right|\right]. \tag{12}$$

We now bound the expectation of the maximum. For each fixed j, the random variable

$$v_j = \frac{1}{m}\sum_{\ell=1}^{m}\varepsilon_\ell\,\phi_j(x_\ell)$$

is a sum of independent, mean-zero, B-bounded terms, hence sub-Gaussian with parameter B^2/m. By Hoeffding's inequality, for any $t > 0$,

$$\Pr_\varepsilon\big(|v_j| \ge t\big) \;\le\; 2\exp\!\left(-\frac{m\,t^2}{2B^2}\right).$$

A union bound over $j = 1, \dots, M$ gives

$$\Pr_\varepsilon\left(\max_{1\le j\le M}|v_j| \ge t\right) \;\le\; 2M\,\exp\!\left(-\frac{m\,t^2}{2B^2}\right).$$

Integrating this tail bound yields the desired expectation control (standard for sub-Gaussian maxima):

$$\begin{aligned}\mathbb{E}_\varepsilon\left[\max_{1\le j\le M}|v_j|\right] &= \int_0^\infty \Pr_\varepsilon\left(\max_j|v_j| \ge t\right)\,dt \\ &\le \int_0^\infty \min\left\{1,\; 2M\,\exp\!\left(-\frac{m\,t^2}{2B^2}\right)\right\}\,dt \;\le\; B\sqrt{\frac{2\log(2M)}{m}}.\end{aligned}$$

Substituting this bound into (12) gives, for the block,

$$\widehat{\mathfrak{R}}_m(\mathcal{S}) \;\le\; \Lambda\,B\sqrt{\frac{2\log(2M)}{m}}.$$

Applying the same argument to the main-effect block (Λ_1, B_1) and the interaction block (Λ_2, B_2) and summing the two contributions yields

$$\widehat{\mathfrak{R}}_m(\mathcal{S}) \;\le\; C_1\,(\Lambda_1 B_1 + \Lambda_2 B_2)\sqrt{\frac{\log M}{m}},$$

after absorbing numerical constants and $\log(2M)$ into $C_1\,\log M$.

If the base function classes $\{f_i\}$ and $\{f_{ij}\}$ are learned and not a fixed finite set but uniformly Lipschitz with constants L_1, L_2 over an input domain of bounded diameter, their empirical Rademacher complexities scale as $O(L/\sqrt{m})$ by Dudley entropy/covering arguments, contributing the second term $C_2(\Lambda_1 L_1 + \Lambda_2 L_2)/\sqrt{m}$. □

Lemma 4. *Let $a_g, b_g \in [0,1]$, $1 \le g \le G$. Then*

$$\left|\operatorname{Var}_g(a_g) - \operatorname{Var}_g(b_g)\right| \;\le\; \frac{4}{G}\sum_{g=1}^{G} |a_g - b_g|.$$

Proof. Write $\operatorname{Var}_g(a) = \frac{1}{G}\sum_g a_g^2 - \left(\frac{1}{G}\sum_g a_g\right)^2$. Using $|u^2 - v^2| \le |u - v|(|u| + |v|)$ and $|u|, |v| \le 1$,

$$\left|\frac{1}{G}\sum_g a_g^2 - \frac{1}{G}\sum_g b_g^2\right| \le \frac{2}{G}\sum_g |a_g - b_g|.$$

Additionally,

$$\left|\left(\tfrac{1}{G}\sum_g a_g\right)^2 - \left(\tfrac{1}{G}\sum_g b_g\right)^2\right| = |(\overline{a} - \overline{b})(\overline{a} + \overline{b})| \le 2\,|\overline{a} - \overline{b}| \le \frac{2}{G}\sum_g |a_g - b_g|.$$

Adding the two previous results gives:

$$\begin{aligned}
\left|\operatorname{Var}_g(a_g) - \operatorname{Var}_g(b_g)\right| &= \left|\frac{1}{G}\sum_g a_g^2 - \left(\tfrac{1}{G}\sum_g a_g\right)^2 - \left(\frac{1}{G}\sum_g b_g^2 - \left(\tfrac{1}{G}\sum_g b_g\right)^2\right)\right| \\
&\le \left|\frac{1}{G}\sum_g a_g^2 - \frac{1}{G}\sum_g b_g^2\right| + \left|\left(\tfrac{1}{G}\sum_g a_g\right)^2 - \left(\tfrac{1}{G}\sum_g b_g\right)^2\right| \\
&\le \frac{2}{G}\sum_g |a_g - b_g| + \frac{2}{G}\sum_g |a_g - b_g| \\
&= \frac{4}{G}\sum_{g=1}^{G} |a_g - b_g|.
\end{aligned}$$

□

Theorem 3 (Alternative formulation of Proposition 3)**.** *Under (11), there exists a constant $C > 0$ such that, with probability at least $1 - \delta$,*

$$\sup_{\theta\in\Theta}\left|\mathcal{P}_{\text{mean}}(\theta) - \widehat{\mathcal{P}}_{\text{mean}}(\theta)\right| \;\le\; \frac{C}{\sqrt{p_{\min}}}\left[(\Lambda_1 B_1 + \Lambda_2 B_2)\sqrt{\frac{\log M}{n}} + (\Lambda_1 L_1 + \Lambda_2 L_2)\,\frac{1}{\sqrt{n}} + \sqrt{\frac{\log(G/\delta)}{n}}\right].$$

Proof. By Lemma 4, for each $y \in \{0,1\}$,

$$\left|\mathrm{Var}_g(\mu_{g,y}) - \mathrm{Var}_g(\widehat{\mu}_{g,y})\right| \;\le\; \frac{4}{G}\sum_{g=1}^{G}\left|\mu_{g,y} - \widehat{\mu}_{g,y}\right| \;\le\; 4\max_{(g,y)}\left|\mu_{g,y} - \widehat{\mu}_{g,y}\right|.$$

Summing over the two labels y gives

$$\sup_{\theta}\left|\mathcal{P}_{\mathrm{mean}}(\theta) - \widehat{\mathcal{P}}_{\mathrm{mean}}(\theta)\right| \;\le\; 8\max_{(g,y)}\sup_{\theta}\left|\mu_{g,y}(\theta) - \widehat{\mu}_{g,y}(\theta)\right|.$$

By applying Lemma 1 to each cell we get, with probability $\ge 1-\delta$,

$$\max_{(g,y)}\sup_{\theta}\left|\mu_{g,y}(\theta) - \widehat{\mu}_{g,y}(\theta)\right| \;\le\; c_1\max_{(g,y)}\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H}) + c_2\max_{(g,y)}\sqrt{\frac{\log(2G/\delta)}{n_{g,y}}}.$$

By Lemma 2 and Lemma 3, we write

$$\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{H}) \;\le\; \frac{1}{4}\,\widehat{\mathfrak{R}}_{n_{g,y}}(\mathcal{S}) \;\le\; \frac{1}{4}\left[C_1(\Lambda_1 B_1 + \Lambda_2 B_2)\sqrt{\frac{\log M}{n_{g,y}}} + C_2(\Lambda_1 L_1 + \Lambda_2 L_2)\frac{1}{\sqrt{n_{g,y}}}\right].$$

For each (g,y), define indicators $Z_\ell^{(g,y)} := \mathbf{1}\{(A_\ell, Y_\ell) = (g,y)\}$ so that $n_{g,y} = \sum_{\ell=1}^{n} Z_\ell^{(g,y)} \sim$ Binomial$(n, p_{g,y})$ with $p_{g,y} = \Pr(A{=}g, Y{=}y) \ge p_{\min}$ by (11). The multiplicative Chernoff bound gives, for any $\eta \in (0,1)$,

$$\Pr\big(n_{g,y} \le (1-\eta)\,n\,p_{g,y}\big) \;\le\; \exp\!\left(-\frac{\eta^2}{2}\,n\,p_{g,y}\right).$$

Taking $\eta = \frac{1}{2}$ and using $p_{g,y} \ge p_{\min}$,

$$\Pr\big(n_{g,y} < \tfrac{1}{2}\,n\,p_{g,y}\big) \;\le\; \exp\!\Big(-\tfrac{1}{8}\,n\,p_{g,y}\Big) \;\le\; \exp\!\Big(-\tfrac{1}{8}\,n\,p_{\min}\Big).$$

A union bound over the $2G$ cells (g,y) yields

$$\Pr\!\left(\min_{(g,y)} n_{g,y} \;\ge\; \tfrac{1}{2}\,n\,p_{\min}\right) \;\ge\; 1 \;-\; 2G\,\exp\!\Big(-\tfrac{1}{8}\,n\,p_{\min}\Big) \;:=\; 1-\delta'.$$

On the event

$$\mathcal{E} := \left\{\min_{(g,y)} n_{g,y} \ge \tfrac{1}{2}\,p_{\min}\,n\right\},$$

we have, for every (g, y),

$$\sqrt{\frac{1}{n_{g,y}}} \le \sqrt{\frac{2}{p_{\min} n}}, \qquad \sqrt{\frac{\log M}{n_{g,y}}} \le \sqrt{\frac{2}{p_{\min}}} \sqrt{\frac{\log M}{n}}, \qquad \sqrt{\frac{\log(2G/\delta)}{n_{g,y}}} \le \sqrt{\frac{2}{p_{\min}}} \sqrt{\frac{\log(2G/\delta)}{n}}.$$

Substituting these bounds into the deviation bound for $\max_{(g,y)} \sup_\theta |\mu_{g,y}(\theta) - \widehat{\mu}_{g,y}(\theta)|$, and then into

$$\sup_\theta \left|\mathcal{P}_{\text{mean}}(\theta) - \widehat{\mathcal{P}}_{\text{mean}}(\theta)\right| \le 8 \max_{(g,y)} \sup_\theta \left|\mu_{g,y}(\theta) - \widehat{\mu}_{g,y}(\theta)\right|,$$

yields

$$\sup_\theta \left|\mathcal{P}_{\text{mean}}(\theta) - \widehat{\mathcal{P}}_{\text{mean}}(\theta)\right| \le \frac{C}{\sqrt{p_{\min}}} \left[(\Lambda_1 B_1 + \Lambda_2 B_2) \sqrt{\frac{\log M}{n}} + (\Lambda_1 L_1 + \Lambda_2 L_2) \frac{1}{\sqrt{n}} + \sqrt{\frac{\log(G/\delta)}{n}} \right],$$

for some absolute constant $C > 0$, on the event $\mathcal{E}$ and the $(1 - \delta)$ event of Lemma 1. By choosing n large enough so that $2G \exp(-np_{\min}/8) \le \delta/2$ (or by absorbing this term into δ), we obtain the stated high-probability bound $1 - \delta$.

□

F Empirical evaluation of penalty

Our training objective augments predictive fit with a soft, threshold-free fairness term that aligns class-conditional scores across protected groups. In Section 2.2, we linked the empirical mean penalty to the weighted-threshold EO disparity bound and show that it generalizes beyond the training set. To verify that the in-sample gains are meaningful, stable, and decision-relevant, we report fairness diagnostics that connect our soft penalty to equalized-odds dispersion near the decision boundary, compute the implemented penalty $\mathcal{P}_{\text{mean}}$ over several subsets, and show learning curves that reflect out-of-sample concentration of these fairness quantities as sample size grows. Together, these exercises assess how the penalty terms behave and whether fairness improvements generalize.

Together with the penalty, we calculate dispersion terms $\mathcal{D}_y$ over a set of thresholds $\mathcal{T}_{\text{eff}}(y)|$, empirically verifying Proposition 2. Monotonic declines of $\mathcal{P}_{\text{mean}}$ and $\mathcal{D}_1$+$\mathcal{D}_0$ in λ across train, validation, and test set, which are shown in Table 10, indicate that in-processing regularization successfully reduces cross-group score shifts within labels and EO dispersion across decision boundaries. We also examine the learning curves for different λ values: To study out-of-sample concentration, we fix each trained model and, for a sequence of sample sizes $n \in \mathcal{N}$ (e.g., n increasing from a small fraction up to the full test set), draw repeated subsamples (without replacement) from the held-out evaluation cohort. On each subsample we recompute $\mathcal{P}_{\text{mean}}$ and $\mathcal{D}_1$+$\mathcal{D}_0$, then plot the mean and ±1 standard deviation across draws. As Figure 9 confirms, under standard regularity, the bands shrink at the usual $\sqrt{n}$ rate, indicating generalization of the fairness results as stated in Proposition 3.

	Model	$\mathcal{P}_{\text{mean}}$	$\mathcal{D}_1$	$\mathcal{D}_0$	$\mathcal{D}_1 + \mathcal{D}_0$	$\lvert\mathcal{T}_{\text{eff}}(y{=}1)\rvert$	$\lvert\mathcal{T}_{\text{eff}}(y{=}0)\rvert$	$n_{A=0,Y=1}$	$n_{A=1,Y=1}$	$n_{A=0,Y=0}$	$n_{A=1,Y=0}$
Train	FAN ($\lambda = 0$)	0.00068	0.00063	0.00044	0.00107	201	201	56639	23291	216022	54048
Train	FAN ($\lambda = 1$)	0.00046	0.00044	0.00027	0.00072	201	201	56639	23291	216022	54048
Train	FAN ($\lambda = 2$)	0.00033	0.00033	0.00018	0.00052	201	201	56639	23291	216022	54048
Train	FAN ($\lambda = 3$)	0.00025	0.00025	0.00016	0.00041	201	201	56639	23291	216022	54048
Validation	FAN ($\lambda = 0$)	0.00058	0.00053	0.00043	0.00097	201	201	12110	5018	46275	11597
Validation	FAN ($\lambda = 1$)	0.00038	0.00037	0.00027	0.00064	201	201	12110	5018	46275	11597
Validation	FAN ($\lambda = 2$)	0.00025	0.00026	0.00019	0.00045	201	201	12110	5018	46275	11597
Validation	FAN ($\lambda = 3$)	0.00019	0.00020	0.00016	0.00036	201	201	12110	5018	46275	11597
Test	FAN ($\lambda = 0$)	0.00079	0.00075	0.00039	0.00114	201	201	12152	4975	46241	11632
Test	FAN ($\lambda = 1$)	0.00055	0.00055	0.00024	0.00079	201	201	12152	4975	46241	11632
Test	FAN ($\lambda = 2$)	0.00042	0.00044	0.00016	0.00060	201	201	12152	4975	46241	11632
Test	FAN ($\lambda = 3$)	0.00031	0.00033	0.00014	0.00046	201	201	12152	4975	46241	11632

Table 10: Penalty term and average (threshold-weighted) group dispersion. This table evaluates penalty and dispersion terms across data sets and models.

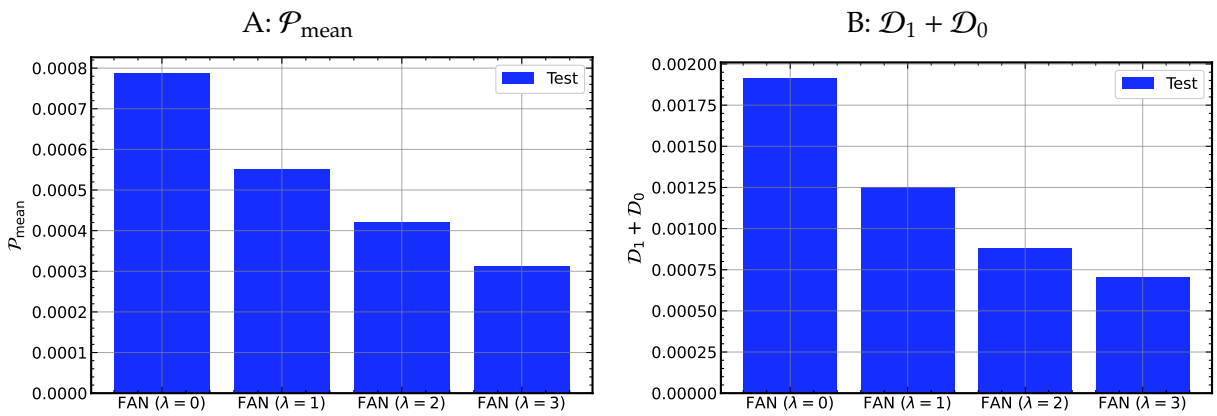


Figure 8: Empirical penalty and upper bound. This figure summarizes the out-of-sample mean penalty (Panel A) and total dispersion bound (Panel B) across penalization levels.

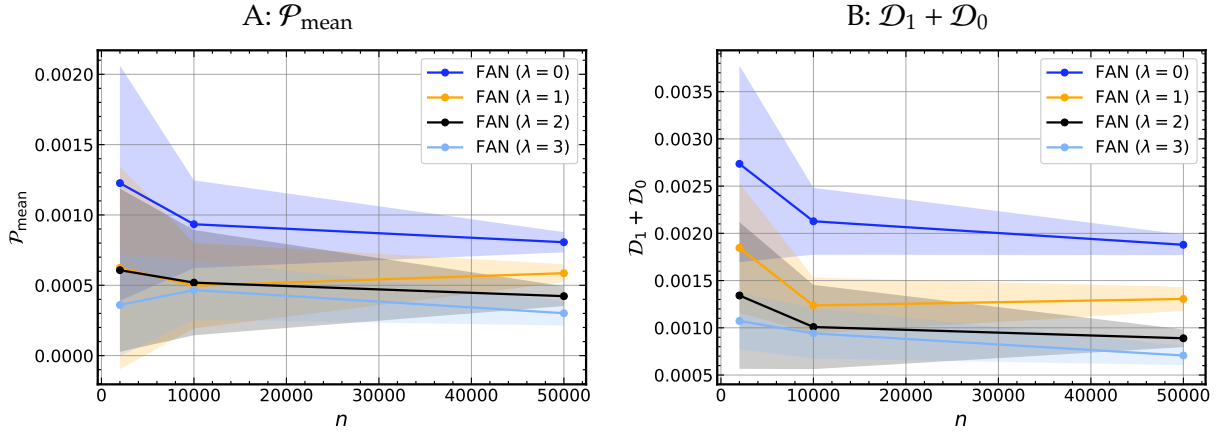


Figure 9: Learning curves. This table reports the mean fairness penalty (Panel A) and dispersion boundary (Panel B) computed over hold-out test sets of varying sizes n. Shaded regions indicate the one standard deviation bands around the mean.

G Robustness: Threshold variation

To assess robustness, we vary the decision threshold τ and, for each model, report the induced share of non-rejections ("approval," since $y = 1$ denotes rejection), overall and group-conditional discrimination (AUC, AUC_0, AUC_1), and disparities in error rates (ΔTPR, ΔFPR), showing that our fairness-accuracy conclusions do not depend on any single arbitrary cutoff.

Recall that each FAN model outputs a score $s(X) \in [0, 1]$. We predict a rejection when $\hat{y}(\tau) = \mathbb{I}\{s(X) \geq \tau\} = 1$ and a non-rejection/approval when $\hat{y}(\tau) = 0$. Note that our target encoding uses $y = 1$ for rejection and $y = 0$ for non-rejection. Therefore the true positive rate (TPR) and false positive rate (FPR) are defined with respect to the rejection class:

$$\text{TPR}(\tau) = \Pr(\hat{y}(\tau) = 1 \mid y = 1), \qquad \text{FPR}(\tau) = \Pr(\hat{y}(\tau) = 1 \mid y = 0).$$

We report the overall ROC AUC, group-conditional AUCs AUC_0 and AUC_1, and the Brier score $\frac{1}{n}\sum_i (s(X_i) - y_i)^2$. Let $A \in \{0, 1\}$ denote group membership. For group-conditional error rates we use $\text{TPR}_a(\tau) = \Pr(\hat{y}(\tau) = 1 \mid y = 1, A = a)$ and $\text{FPR}_a(\tau) = \Pr(\hat{y}(\tau) = 1 \mid y = 0, A = a)$, and report disparity measures

$$\Delta\text{TPR}(\tau) = \text{TPR}_0(\tau) - \text{TPR}_1(\tau), \quad \Delta\text{FPR}(\tau) = \text{FPR}_0(\tau) - \text{FPR}_1(\tau),$$

together with their absolute values $|\Delta\text{TPR}(\tau)|$ and $|\Delta\text{FPR}(\tau)|$. We refer to the approval rate as $\alpha(\tau) = \Pr(\hat{y}(\tau) = 0)$, i.e., the fraction of instances not flagged for rejection.

To make model comparisons we hold the overall decision rate fixed and select a single decision threshold $\tau^\star$ per model such that the model's approval share matches the empirical approval share in the test data,

$$\alpha(\tau^\star) = \alpha_{\text{obs}} \quad \text{with} \quad \alpha_{\text{obs}} = 1 - \frac{1}{n_{\text{test}}} \sum_{i \in \mathcal{T}_{\text{test}}} y_i.$$

Table 11 summarizes test-set performance for FAN models with $\lambda \in \{0, 1, 2, 3\}$ at these operating thresholds $\tau^\star$ that equate the model's approval share with the observed test-set approval share. Across λ, overall AUC remains stable (0.814 - 0.816), while increasing λ reduces the group disparities at $\tau^\star$ (both $|\Delta\text{TPR}|$ and $|\Delta\text{FPR}|$ decrease monotonically), with modest trade-offs in the Brier score.

For the models with $\lambda_{EO} = 0$ and $\lambda_{EO} = 3$ we report a threshold sweep in Tables 12 and 13, respectively, showing how the approval rate $\alpha(\tau)$, group-conditional error rates $\text{TPR}_a(\tau)$ and $\text{FPR}_a(\tau)$, and disparity measures $\Delta\text{TPR}(\tau)$ and $\Delta\text{FPR}(\tau)$ evolve with the

decision threshold τ. Because $y = 1$ encodes rejection, higher τ corresponds to fewer predicted rejections and thus a larger approval share.

Model	$\tau^\star$	AUC	AUC_0	AUC_1	$\Delta TPR(\tau^\star)$	$\lvert\Delta TPR(\tau^\star)\rvert$	$\Delta FPR(\tau^\star)$	$\lvert\Delta FPR(\tau^\star)\rvert$	Brier
FAN ($\lambda = 0$)	0.2776	0.8234	0.8200	0.8243	0.0768	0.0768	0.0428	0.0428	0.1216
FAN ($\lambda = 1$)	0.2780	0.8233	0.8201	0.8249	0.0604	0.0604	0.0310	0.0310	0.1217
FAN ($\lambda = 2$)	0.2805	0.8242	0.8215	0.8255	0.0529	0.0529	0.0238	0.0238	0.1212
FAN ($\lambda = 3$)	0.2854	0.8237	0.8212	0.8251	0.0391	0.0391	0.0159	0.0159	0.1214

Table 11: Fairness summary for observed rate-matched decision thresholds. This table reports performance and fairness metrics across models with matched decision thresholds.

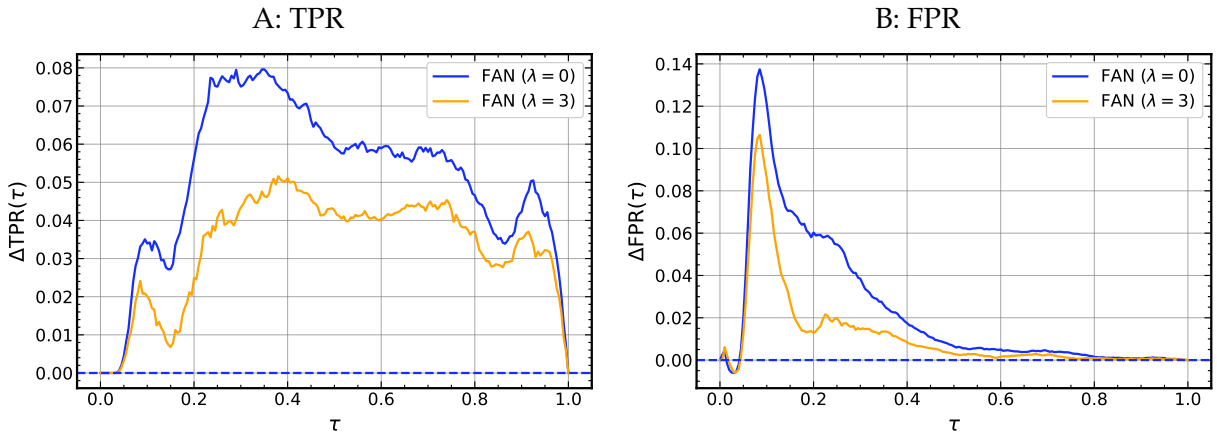


Figure 10: Rate disparities across thresholds. This figure presents group differences in true positive rates (Panel A) and false positive rates (Panel B) across decision thresholds, comparing the unconstrained and the penalized classifier.

τ	Appr. rate	TPR_0	TPR_1	ΔTPR	$\lvert\Delta\text{TPR}\rvert$	FPR_0	FPR_1	ΔFPR	$\lvert\Delta\text{FPR}\rvert$
0.0099	0.0200	1.0000	1.0000	0.0000	0.0000	0.9731	0.9781	0.0050	0.0050
0.0466	0.0446	0.9968	0.9996	0.0028	0.0028	0.9410	0.9504	0.0094	0.0094
0.0532	0.0692	0.9919	0.9980	0.0061	0.0061	0.9054	0.9389	0.0334	0.0334
0.0579	0.0939	0.9861	0.9962	0.0101	0.0101	0.8709	0.9244	0.0536	0.0536
0.0617	0.1185	0.9797	0.9940	0.0143	0.0143	0.8363	0.9106	0.0743	0.0743
0.0653	0.1431	0.9741	0.9922	0.0181	0.0181	0.8021	0.8947	0.0926	0.0926
0.0688	0.1677	0.9670	0.9885	0.0215	0.0215	0.7695	0.8746	0.1051	0.1051
0.0724	0.1923	0.9589	0.9839	0.0250	0.0250	0.7371	0.8551	0.1181	0.1181
0.0761	0.2169	0.9517	0.9799	0.0282	0.0282	0.7051	0.8327	0.1276	0.1276
0.0800	0.2415	0.9440	0.9741	0.0301	0.0301	0.6740	0.8080	0.1340	0.1340
0.0840	0.2662	0.9351	0.9684	0.0334	0.0334	0.6440	0.7806	0.1367	0.1367
0.0884	0.2908	0.9255	0.9598	0.0343	0.0343	0.6152	0.7498	0.1346	0.1346
0.0935	0.3154	0.9165	0.9520	0.0355	0.0355	0.5869	0.7165	0.1296	0.1296
0.0988	0.3400	0.9071	0.9421	0.0350	0.0350	0.5590	0.6828	0.1238	0.1238
0.1052	0.3646	0.8971	0.9311	0.0340	0.0340	0.5326	0.6443	0.1117	0.1117
0.1125	0.3892	0.8851	0.9184	0.0333	0.0333	0.5067	0.6063	0.0997	0.0997
0.1204	0.4139	0.8738	0.9071	0.0333	0.0333	0.4805	0.5684	0.0880	0.0880
0.1290	0.4385	0.8617	0.8917	0.0300	0.0300	0.4541	0.5340	0.0800	0.0800
0.1382	0.4631	0.8477	0.8764	0.0287	0.0287	0.4275	0.5021	0.0746	0.0746
0.1471	0.4877	0.8344	0.8605	0.0261	0.0261	0.4004	0.4719	0.0715	0.0715
0.1562	0.5123	0.8175	0.8466	0.0292	0.0292	0.3737	0.4427	0.0690	0.0690
0.1651	0.5369	0.7999	0.8354	0.0355	0.0355	0.3467	0.4145	0.0678	0.0678
0.1739	0.5615	0.7811	0.8195	0.0384	0.0384	0.3209	0.3847	0.0638	0.0638
0.1831	0.5861	0.7605	0.8058	0.0453	0.0453	0.2948	0.3573	0.0625	0.0625
0.1923	0.6108	0.7400	0.7912	0.0511	0.0511	0.2690	0.3288	0.0599	0.0599
0.2020	0.6354	0.7162	0.7733	0.0571	0.0571	0.2437	0.3034	0.0597	0.0597
0.2128	0.6600	0.6952	0.7576	0.0624	0.0624	0.2177	0.2766	0.0589	0.0589
0.2241	0.6846	0.6699	0.7377	0.0678	0.0678	0.1930	0.2509	0.0578	0.0578
0.2370	0.7092	0.6405	0.7182	0.0777	0.0777	0.1689	0.2270	0.0581	0.0581
0.2509	0.7338	0.6142	0.6890	0.0748	0.0748	0.1458	0.2001	0.0542	0.0542
0.2676	0.7585	0.5851	0.6621	0.0770	0.0770	0.1233	0.1730	0.0497	0.0497
0.2875	0.7831	0.5524	0.6314	0.0789	0.0789	0.1026	0.1437	0.0410	0.0410
0.3140	0.8077	0.5202	0.5964	0.0762	0.0762	0.0821	0.1152	0.0331	0.0331
0.3475	0.8323	0.4816	0.5608	0.0792	0.0792	0.0627	0.0892	0.0265	0.0265
0.3919	0.8569	0.4366	0.5114	0.0747	0.0747	0.0460	0.0648	0.0188	0.0188
0.4675	0.8815	0.3896	0.4545	0.0649	0.0649	0.0308	0.0400	0.0092	0.0092
0.5949	0.9061	0.3266	0.3857	0.0591	0.0591	0.0189	0.0237	0.0048	0.0048
0.7412	0.9308	0.2534	0.3089	0.0556	0.0556	0.0092	0.0129	0.0037	0.0037
0.8968	0.9554	0.1699	0.2123	0.0423	0.0423	0.0038	0.0045	0.0007	0.0007
0.9512	0.9800	0.0721	0.1140	0.0419	0.0419	0.0008	0.0016	0.0008	0.0008

Table 12: Metrics at varying thresholds ($\lambda_{EO} = 0$)

τ	Appr. rate	TPR_0	TPR_1	ΔTPR	$\lvert\Delta$TPR$\rvert$	FPR_0	FPR_1	ΔFPR	$\lvert\Delta$FPR$\rvert$
0.0149	0.0200	1.0000	1.0000	0.0000	0.0000	0.9736	0.9760	0.0024	0.0024
0.0488	0.0446	0.9971	0.9996	0.0025	0.0025	0.9410	0.9501	0.0091	0.0091
0.0559	0.0692	0.9916	0.9976	0.0060	0.0060	0.9061	0.9365	0.0303	0.0303
0.0605	0.0939	0.9862	0.9952	0.0090	0.0090	0.8719	0.9206	0.0488	0.0488
0.0642	0.1185	0.9806	0.9908	0.0102	0.0102	0.8390	0.9004	0.0615	0.0615
0.0678	0.1431	0.9744	0.9871	0.0127	0.0127	0.8060	0.8811	0.0752	0.0752
0.0711	0.1677	0.9677	0.9841	0.0164	0.0164	0.7732	0.8608	0.0876	0.0876
0.0745	0.1923	0.9610	0.9791	0.0181	0.0181	0.7414	0.8379	0.0966	0.0966
0.0782	0.2169	0.9533	0.9739	0.0206	0.0206	0.7103	0.8128	0.1025	0.1025
0.0817	0.2415	0.9458	0.9674	0.0217	0.0217	0.6802	0.7844	0.1042	0.1042
0.0858	0.2662	0.9377	0.9612	0.0235	0.0235	0.6505	0.7551	0.1046	0.1046
0.0899	0.2908	0.9291	0.9497	0.0207	0.0207	0.6222	0.7226	0.1004	0.1004
0.0946	0.3154	0.9203	0.9411	0.0208	0.0208	0.5941	0.6884	0.0943	0.0943
0.1000	0.3400	0.9116	0.9317	0.0200	0.0200	0.5665	0.6528	0.0863	0.0863
0.1061	0.3646	0.9021	0.9202	0.0181	0.0181	0.5396	0.6156	0.0760	0.0760
0.1133	0.3892	0.8901	0.9069	0.0168	0.0168	0.5139	0.5772	0.0633	0.0633
0.1214	0.4139	0.8784	0.8911	0.0127	0.0127	0.4879	0.5408	0.0529	0.0529
0.1305	0.4385	0.8655	0.8772	0.0117	0.0117	0.4624	0.5030	0.0406	0.0406
0.1400	0.4631	0.8539	0.8625	0.0087	0.0087	0.4352	0.4709	0.0357	0.0357
0.1499	0.4877	0.8392	0.8462	0.0070	0.0070	0.4094	0.4372	0.0279	0.0279
0.1594	0.5123	0.8228	0.8340	0.0111	0.0111	0.3835	0.4035	0.0200	0.0200
0.1691	0.5369	0.8072	0.8179	0.0107	0.0107	0.3570	0.3736	0.0166	0.0166
0.1791	0.5615	0.7883	0.8050	0.0167	0.0167	0.3309	0.3436	0.0127	0.0127
0.1889	0.5861	0.7704	0.7912	0.0207	0.0207	0.3040	0.3167	0.0127	0.0127
0.1988	0.6108	0.7498	0.7737	0.0239	0.0239	0.2777	0.2914	0.0137	0.0137
0.2094	0.6354	0.7281	0.7554	0.0273	0.0273	0.2518	0.2662	0.0145	0.0145
0.2204	0.6600	0.7033	0.7391	0.0358	0.0358	0.2254	0.2454	0.0201	0.0201
0.2323	0.6846	0.6796	0.7154	0.0358	0.0358	0.2007	0.2199	0.0192	0.0192
0.2455	0.7092	0.6542	0.6915	0.0372	0.0372	0.1761	0.1958	0.0197	0.0197
0.2598	0.7338	0.6250	0.6681	0.0431	0.0431	0.1530	0.1694	0.0164	0.0164
0.2758	0.7585	0.5970	0.6370	0.0400	0.0400	0.1296	0.1461	0.0166	0.0166
0.2949	0.7831	0.5638	0.6054	0.0417	0.0417	0.1077	0.1228	0.0151	0.0151
0.3204	0.8077	0.5287	0.5757	0.0470	0.0470	0.0857	0.1005	0.0148	0.0148
0.3548	0.8323	0.4914	0.5375	0.0460	0.0460	0.0652	0.0787	0.0134	0.0134
0.3981	0.8569	0.4443	0.4943	0.0500	0.0500	0.0479	0.0564	0.0085	0.0085
0.4728	0.8815	0.3957	0.4374	0.0416	0.0416	0.0319	0.0365	0.0045	0.0045
0.6005	0.9061	0.3320	0.3727	0.0406	0.0406	0.0195	0.0211	0.0016	0.0016
0.7354	0.9308	0.2567	0.3011	0.0444	0.0444	0.0097	0.0110	0.0013	0.0013
0.8905	0.9554	0.1724	0.2062	0.0338	0.0338	0.0038	0.0041	0.0003	0.0003
0.9472	0.9800	0.0751	0.1079	0.0328	0.0328	0.0008	0.0013	0.0005	0.0005

Table 13: Metrics at varying thresholds ($\lambda_{EO} = 3$)

H Robustness: Alternative cutoffs

H.1 Threshold analysis using DTI cutoff

For robustness, we implement the same local RD around the underwriting boundary at DTI≈43%, which served as the "General QM" limit under the Ability-to-Repay/Qualified Mortgage regime during our sample period.[30] Let $r_l := \text{DTI}_l - 43$ (in percentage points) be the running variable and $D_l := \mathbb{1}\{r_l \geq 0\}$ the indicator for being (weakly) above the threshold. We estimate the local-linear specification in Eq. (4) with triangular kernel weights $w_l(h) = \max\{0, 1 - |r_l|/h\}$ on $|r_l| \leq h$, via WLS with heteroskedasticity-consistent standard errors. As before, γ is the non-minority jump at the cutoff and ζ is the minority differential, so the minority jump equals $\gamma+\zeta$.

Table 14 reports our main specification with a bandwidth of h=10 p.p. Model-generated approvals exhibit a sizable and precisely estimated negative discontinuity at DTI≈ 43% for non-minority applicants: $\hat{\gamma}$ ranges from −7.41 to −6.39 p.p. across λ (all p<0.001). The minority-specific differential is negative and statistically significant but somewhat smaller in magnitude than at the LTV boundary, with $\hat{\zeta}$ between −3.17 and −2.58 p.p. (all p<0.001). Observed approvals in HMDA display the same sign pattern with a smaller baseline jump, $\hat{\gamma}$ = −4.03 p.p. (p<0.001), and a minority differential of −2.63 p.p. (p<0.001).

Figure 11 presents local linear fits on each side of the cutoff, illustrating the discontinuity.[31] Table 15 shows that estimates vary somewhat across bandwidths, consistent with local curvature: For model approvals, the estimated jump $\hat{\gamma}$ is near zero or slightly positive at the narrowest window (h=3), becomes close to zero to mildly negative at h=5, and is clearly negative and precisely estimated at h=10. Crucially, the minority differential $\hat{\zeta}$ is negative, stable in sign, and statistically significant across all bandwidths and all fairness weights. Moreover, $\hat{\zeta}$ declines monotonically as the fairness penalty increases at each bandwidth: for example, at h=10 it shrinks from −3.17 p.p. (λ=0) to −2.58 p.p. (λ=3), with analogous decreases at h=3 and h=5. Observed approvals mirror these patterns: $\hat{\gamma}$ shifts from slightly positive at h=3 to negative by h=10, while $\hat{\zeta}$ remains negative and precisely estimated across bandwidths (between roughly −3.43 and −2.63 p.p.).

Overall, this confirms our finding in Section 4 that our fairness regularization successfully reduces minority-specific approval gaps at the decision boundary.

[30] DeFusco et al. (2020) establish a discontinuity at the cutoff, showing higher rates at the threshold and a contraction in high-DTI originations with high concentration just below.

[31] Note that the scatter plot represents means within equally spaced bins, so bins have different numbers of observations influencing the linear estimate.

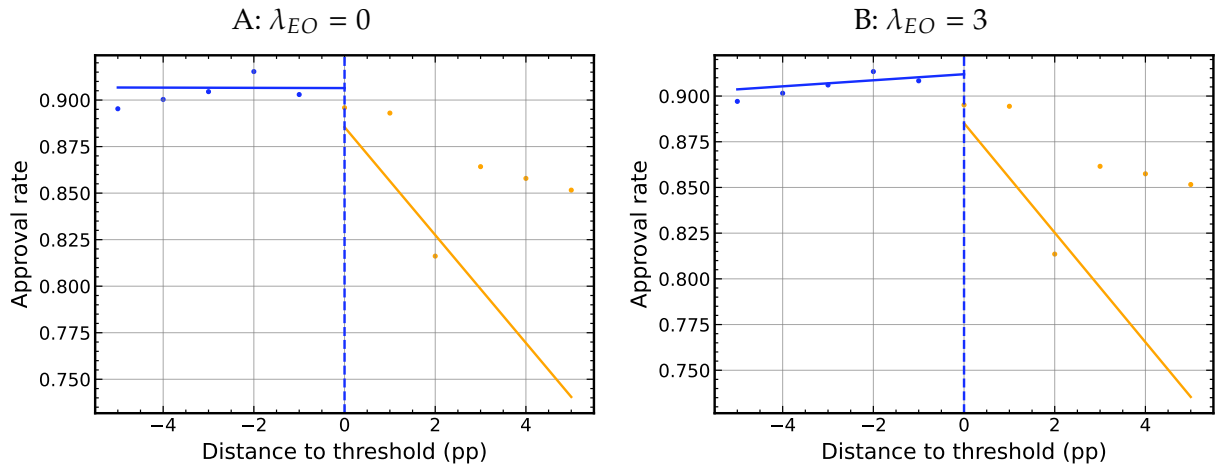


Figure 11: Binned means and local linear fit at DTI=43% for model approvals (h=5). The figure shows loan-to-value ratios and approval rates for the unconstrained FAN (Panel A) and the penalized model (Panel B) with the corresponding regression lines below and above the 80% threshold.

Outcome	N	h	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\zeta}$
Model approvals, λ=0	63886	10.0	0.9224***	-0.0639***	-0.0317***
			(0.0050)	(0.0059)	(0.0081)
Model approvals, λ=1	63886	10.0	0.9226***	-0.0687***	-0.0305***
			(0.0050)	(0.0059)	(0.0079)
Model approvals, λ=2	63886	10.0	0.9180***	-0.0741***	-0.0268***
			(0.0050)	(0.0060)	(0.0079)
Model approvals, λ=3	63886	10.0	0.9194***	-0.0698***	-0.0258***
			(0.0050)	(0.0060)	(0.0077)
Observed approvals	426396	10.0	0.8738***	-0.0403***	-0.0263***
			(0.0023)	(0.0027)	(0.0035)

Table 14: Regression results around DTI= 43% (bandwidth h=10 p.p.). This table reports intercepts and discontinuities estimated from the RD design specified in (4). N is the size of the estimation sample taken from the test data for model approvals and from the full data for observed approvals. Significance levels are denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

h	N	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\zeta}$
		Model approvals, λ=0		
3.0	46924	0.9006***	0.0318**	-0.0562***
		(0.0123)	(0.0134)	(0.0122)
5.0	56344	0.9185***	-0.0097	-0.0483***
		(0.0072)	(0.0086)	(0.0098)
10.0	63886	0.9224***	-0.0639***	-0.0317***
		(0.0050)	(0.0059)	(0.0081)
		Model approvals, λ=1		
3.0	46924	0.9030***	0.0239*	-0.0562***
		(0.0122)	(0.0134)	(0.0117)
5.0	56344	0.9188***	-0.0150*	-0.0473***
		(0.0072)	(0.0086)	(0.0095)
10.0	63886	0.9226***	-0.0687***	-0.0305***
		(0.0050)	(0.0059)	(0.0079)
		Model approvals, λ=2		
3.0	46924	0.9021***	0.0250*	-0.0514***
		(0.0124)	(0.0136)	(0.0117)
5.0	56344	0.9149***	-0.0150*	-0.0421***
		(0.0073)	(0.0087)	(0.0095)
10.0	63886	0.9180***	-0.0741***	-0.0268***
		(0.0050)	(0.0060)	(0.0079)
		Model approvals, λ=3		
3.0	46924	0.9066***	0.0168	-0.0486***
		(0.0122)	(0.0134)	(0.0113)
5.0	56344	0.9172***	-0.0171**	-0.0414***
		(0.0072)	(0.0086)	(0.0092)
10.0	63886	0.9194***	-0.0698***	-0.0258***
		(0.0050)	(0.0060)	(0.0077)
		Observed approvals		
3.0	312870	0.8621***	0.0124**	-0.0313***
		(0.0057)	(0.0062)	(0.0055)
5.0	376639	0.8700***	-0.0071*	-0.0343***
		(0.0033)	(0.0040)	(0.0043)
10.0	426396	0.8738***	-0.0403***	-0.0263***
		(0.0023)	(0.0027)	(0.0035)

Table 15: Regression results around DTI=43% across bandwidths $h \in \{3, 5, 10\}$. This table reports intercepts and discontinuities estimated from the RD design specified in (4) across bandwidths. N is the size of the estimation sample taken from the test data for model approvals and from the full data for observed approvals. Significance levels are denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

H.2 Placebo test

To demonstrate the validity of our RD design, we repeat the analysis using a placebo cutoff of LTV=70%. If our design is well specified, approval probabilities should vary smoothly around this placebo threshold, yielding no statistically meaningful jumps for either group and no systematic minority differential. Table 16 reports estimates for model-generated approvals across bandwidths $h \in \{3, 5, 10\}$ and fairness weights $\lambda \in \{0, 1, 2, 3\}$. The estimated jump at the threshold for non-minority applicants, $\hat{\gamma}$, is small and statistically indistinguishable from zero in nearly all specifications. The only exception is a borderline positive, weakly significant estimate at h=10, λ=2 (1.86 p.p.), which is not replicated at neighboring λ or h, consistent with sampling variation rather than a systematic effect. The minority-specific differential at the cutoff, $\hat{\zeta}$, is also small in magnitude (roughly 0–4 p.p.) and imprecisely estimated across all bandwidths and fairness settings, with no consistent pattern in sign or size. Varying the bandwidth does not reveal any emerging discontinuity: estimates remain close to zero at h=3 and 5, and the h=10 results similarly show no robust departure from continuity.

We draw two main conclusions: First, the absence of significant jumps in approval rates at LTV= 70% supports the main design and indicates that our main findings at LTV= 80% are not artifacts of functional form or local composition around arbitrary thresholds. Second, the placebo results show that fairness regularization does not introduce spurious discontinuities away from binding rules: as λ increases, neither $\hat{\gamma}$ nor $\hat{\zeta}$ displays a systematic shift at the placebo cutoff. In combination with the main RD results, this pattern strengthens the interpretation that the documented decline in the minority-specific discontinuity $\hat{\zeta}$ with higher fairness weights reflects targeted changes rather than global distortions that create new discontinuities elsewhere.

h	N	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\zeta}$
Model approvals, λ=0				
3.0	4423	0.9295***	-0.0007	0.0394
		(0.0130)	(0.0158)	(0.0318)
5.0	8225	0.9172***	0.0096	0.0235
		(0.0105)	(0.0131)	(0.0250)
10.0	27521	0.9133***	0.0143	0.0134
		(0.0080)	(0.0103)	(0.0174)
Model approvals, λ=1				
3.0	4423	0.9298***	-0.0019	0.0360
		(0.0129)	(0.0158)	(0.0316)
5.0	8225	0.9178***	0.0085	0.0214
		(0.0105)	(0.0131)	(0.0249)
10.0	27521	0.9138***	0.0131	0.0114
		(0.0080)	(0.0103)	(0.0174)
Model approvals, λ=2				
3.0	4423	0.9285***	0.0051	0.0308
		(0.0130)	(0.0157)	(0.0321)
5.0	8225	0.9149***	0.0149	0.0188
		(0.0106)	(0.0131)	(0.0254)
10.0	27521	0.9111***	0.0186*	0.0094
		(0.0081)	(0.0103)	(0.0176)
Model approvals, λ=3				
3.0	4423	0.9288***	-0.0020	0.0370
		(0.0131)	(0.0159)	(0.0317)
5.0	8225	0.9152***	0.0098	0.0232
		(0.0106)	(0.0133)	(0.0250)
10.0	27521	0.9117***	0.0140	0.0088
		(0.0081)	(0.0103)	(0.0174)
Observed approvals				
3.0	29660	0.7973***	0.0308***	-0.0121
		(0.0078)	(0.0093)	(0.0162)
5.0	55399	0.7972***	0.0240***	0.0023
		(0.0062)	(0.0076)	(0.0129)
10.0	184031	0.7954***	0.0183***	0.0178*
		(0.0045)	(0.0058)	(0.0091)

Table 16: Regression results around LTV= 70% across bandwidths $h \in \{3, 5, 10\}$. This table reports intercepts and discontinuities estimated from the RD design specified in (4) across bandwidths around the placebo cutoff. N is the size of the estimation sample taken from the test data for model approvals and from the full data for observed approvals. Significance levels are denoted by asterisks: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

H.3 Uniform kernel estimation

While our main RD estimations incorporate triangular kernels, increasing sample weights close to the cutoff, we also re-estimate our RD designs with a uniform kernel. We confirm that our main findings are not sensitive to the choice of weighting scheme, reporting uniform weighting results for both the DTI (Figure 12 and Table 18) and the LTV cutoff (Figure 13 and Table 17).

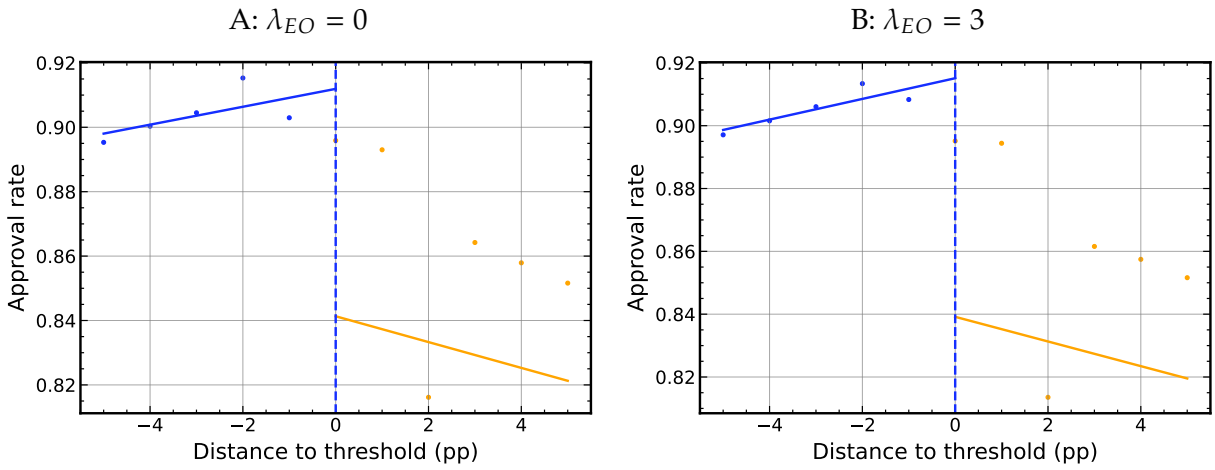


Figure 12: Local linear fit at DTI=43% for model approvals (uniform kernel)

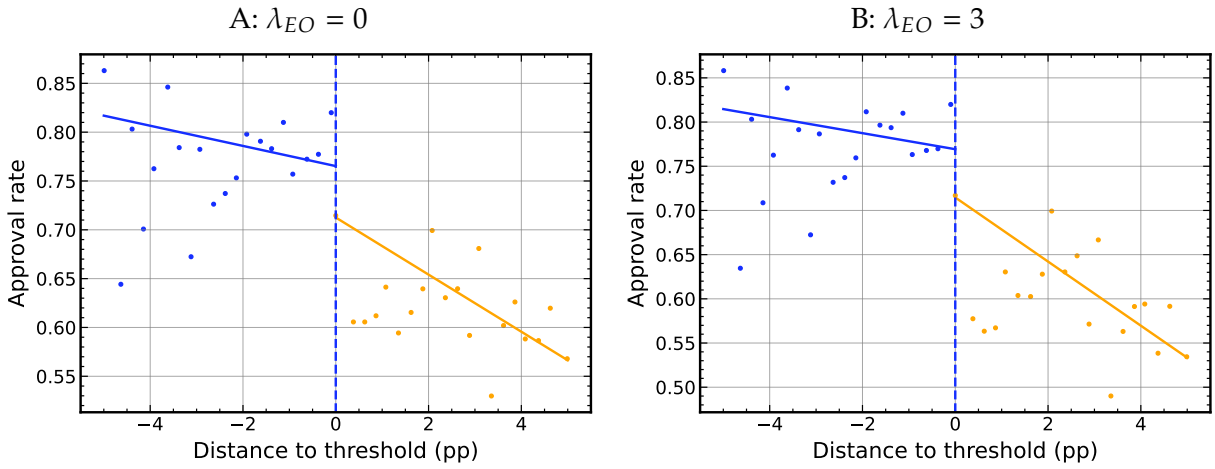


Figure 13: Local linear fit at LTV=80% for model approvals (uniform kernel)

h	N	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\zeta}$
		Model approvals, λ=0		
3.0	16710	0.9241***	-0.0754***	-0.0723***
		(0.0098)	(0.0104)	(0.0198)
5.0	22536	0.9086***	-0.0616***	-0.0707***
		(0.0075)	(0.0082)	(0.0149)
10.0	31463	0.9126***	-0.0687***	-0.0627***
		(0.0059)	(0.0067)	(0.0115)
		Model approvals, λ=1		
3.0	16710	0.9236***	-0.0741***	-0.0668***
		(0.0098)	(0.0104)	(0.0199)
5.0	22536	0.9081***	-0.0606***	-0.0646***
		(0.0075)	(0.0082)	(0.0149)
10.0	31463	0.9113***	-0.0672***	-0.0583***
		(0.0059)	(0.0067)	(0.0115)
		Model approvals, λ=2		
3.0	16710	0.9220***	-0.0581***	-0.0716***
		(0.0097)	(0.0103)	(0.0198)
5.0	22536	0.9071***	-0.0455***	-0.0649***
		(0.0076)	(0.0082)	(0.0149)
10.0	31463	0.9105***	-0.0536***	-0.0556***
		(0.0059)	(0.0067)	(0.0115)
		Model approvals, λ=3		
3.0	16710	0.9223***	-0.0769***	-0.0645***
		(0.0098)	(0.0104)	(0.0198)
5.0	22536	0.9087***	-0.0653***	-0.0608***
		(0.0075)	(0.0082)	(0.0149)
10.0	31463	0.9125***	-0.0720***	-0.0551***
		(0.0059)	(0.0067)	(0.0114)
		Observed approvals		
3.0	111406	0.8004***	-0.0340***	-0.0211**
		(0.0054)	(0.0056)	(0.0099)
5.0	150550	0.7745***	-0.0100**	-0.0323***
		(0.0043)	(0.0045)	(0.0075)
10.0	209977	0.7864***	-0.0387***	-0.0250***
		(0.0033)	(0.0036)	(0.0058)

Table 17: Local linear fit at LTV=80% across bandwidths $h \in \{3, 5, 10\}$

h	N	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\zeta}$
		Model approvals, λ=0		
3.0	46924	0.9167***	-0.0055	-0.0571***
		(0.0085)	(0.0100)	(0.0102)
5.0	56344	0.9270***	-0.0692***	-0.0305***
		(0.0060)	(0.0071)	(0.0088)
10.0	63886	0.9225***	-0.0757***	-0.0262***
		(0.0049)	(0.0059)	(0.0079)
		Model approvals, λ=1		
3.0	46924	0.9174***	-0.0114	-0.0553***
		(0.0085)	(0.0099)	(0.0099)
5.0	56344	0.9257***	-0.0720***	-0.0301***
		(0.0060)	(0.0071)	(0.0085)
10.0	63886	0.9230***	-0.0806***	-0.0245***
		(0.0048)	(0.0059)	(0.0077)
		Model approvals, λ=2		
3.0	46924	0.9168***	-0.0134	-0.0476***
		(0.0086)	(0.0100)	(0.0099)
5.0	56344	0.9201***	-0.0771***	-0.0285***
		(0.0061)	(0.0072)	(0.0085)
10.0	63886	0.9186***	-0.0876***	-0.0213***
		(0.0049)	(0.0059)	(0.0077)
		Model approvals, λ=3		
3.0	46924	0.9157***	-0.0127	-0.0505***
		(0.0084)	(0.0099)	(0.0096)
5.0	56344	0.9230***	-0.0735***	-0.0255***
		(0.0060)	(0.0071)	(0.0083)
10.0	63886	0.9195***	-0.0816***	-0.0204***
		(0.0049)	(0.0059)	(0.0075)
		Observed approvals		
3.0	312870	0.8685***	-0.0037	-0.0401***
		(0.0040)	(0.0046)	(0.0045)
5.0	376639	0.8743***	-0.0397***	-0.0282***
		(0.0028)	(0.0032)	(0.0037)
10.0	426396	0.8744***	-0.0481***	-0.0228***
		(0.0023)	(0.0027)	(0.0033)

Table 18: Local linear fit at DTI=43% across bandwidths $h \in \{3, 5, 10\}$ (uniform kernel)

I Robustness: Removing other sensitive attributes

Our main analysis considers disparities across racial/ethnic groups. However, applicant sex and age may also be viewed as sensitive or potentially inadmissible inputs in a mortgage-screening context. To assess whether our main findings depend on the inclusion of these variables, we re-estimate the FAN models after removing all sex- and age-related predictors from the feature set prior to training. We then repeat the same training, validation, and test procedure used in the baseline analysis.

Table 19 shows that the results are very similar to those reported in the main specification. First, predictive performance changes only minimally. Relative to the main analysis, test accuracy is virtually unchanged for both the Logit and FAN models, while ROC-AUC declines only slightly for the FAN specifications (by about 0.001-0.002 depending on the fairness weight). Second, the fairness patterns remain intact. As in the baseline results, increasing the equalized-odds penalty monotonically reduces the minority-non-minority TPR gap, from 0.058 at $\lambda_{EO} = 0$ to 0.043 at $\lambda_{EO} = 3$. The FPR gap also remains small throughout and declines from 0.007 to 0.003 as the penalty increases.

A comparison with the main table also shows that excluding age and sex does not materially weaken the relative performance of FAN. The unconstrained and penalized FAN models continue to outperform the Logit benchmark in both accuracy and ROC-AUC, while delivering substantially smaller group disparities in both TPR and FPR.

Overall, the evidence suggests that the fairness-performance trade-off documented in the paper is not driven by the direct inclusion of applicant age or sex. Instead, the model continues to deliver monotonic improvements in equalized-odds disparity with negligible changes in out-of-sample fit even when these variables are excluded.

	Accuracy	ROC–AUC	$\lvert TPR_1 - TPR_0 \rvert$	$\lvert FPR_1 - FPR_0 \rvert$
Logit (baseline)	0.829	0.801	0.064	0.012
FAN ($\lambda_{EO} = 0$)	0.838	0.821	0.058	0.007
FAN ($\lambda_{EO} = 1$)	0.839	0.822	0.051	0.004
FAN ($\lambda_{EO} = 2$)	0.839	0.822	0.045	0.004
FAN ($\lambda_{EO} = 3$)	0.839	0.822	0.043	0.003

Table 19: Performance and fairness metrics for baseline Logit and FAN models on the test data without age and sex variables. The fairness metrics are reported as absolute differences in group-specific rates between minority ($a = 1$) and non-minority ($a = 0$).

J Robustness: Alternative data

As a test of robustness, we repeat our analysis on 2019 HMDA data. Table 20 shows that the 2019 results closely mirror (and in several respects strengthen) the main 2018 findings in Table 1. Predictive performance is stable or slightly higher in 2019: the baseline Logit accuracy rises from 0.829 (2018) to 0.861 (2019), and FAN models achieve 0.867–0.868 accuracy in 2019 versus 0.838–0.840 in 2018, with ROC-AUCs essentially unchanged (around 0.822–0.824 for FAN across both years). Crucially, the fairness regularizer continues to deliver monotonic improvements in equalized odds gaps. The absolute TPR gap decreases from 0.059 to 0.043 across λ_{EO}=0 to 3 in 2018 and from 0.047 to 0.028 in 2019, while the absolute FPR gap shrinks from 0.006 to 0.003 in 2018 and from 0.009 to 0.003 in 2019. Therefore, the 2019 exercise confirms that stronger fairness regularization improves both components of EO without material losses in accuracy or AUC.

Feature importance rankings are also remarkably consistent across years and regularization levels (see Figure 14). In both 2018 and 2019, `loan_purpose_1` and `dti_num` remain the dominant drivers at λ_{EO}=0 and λ_{EO}=3, alongside `log1p(loan_amount)` and `loan_to_value_ratio`. As in the main results, increasing λ_{EO} reduces the prominence of geographic/minority proxy variables (e.g., `tract_minority_population_percent` drops out of the top set by λ_{EO}=3) and elevates income- and contract-structure terms and interactions (e.g., `dti_num` × `loan_type_1`, `log1p(income)` and its interactions, and `tract_to_msa_income_percentage`). Furthermore, the shape functions associated with the most important predictors as well as the qualitative patterns observed for the fairness adjustments match the main analysis (see Figure 15). Overall, the 2019 robustness test supports the main conclusions from the original analysis: the FAN approach attains small but consistent gains in predictive performance while delivering clear, monotonic improvements in equalized odds, with stable and interpretable importance structures across years.

	Accuracy	ROC–AUC	$\lvert\text{TPR}_1 - \text{TPR}_0\rvert$	$\lvert\text{FPR}_1 - \text{FPR}_0\rvert$
Logit (baseline)	0.861	0.805	0.047	0.012
FAN ($\lambda_{\text{EO}} = 0$)	0.867	0.824	0.047	0.009
FAN ($\lambda_{\text{EO}} = 1$)	0.868	0.823	0.040	0.007
FAN ($\lambda_{\text{EO}} = 2$)	0.868	0.824	0.033	0.004
FAN ($\lambda_{\text{EO}} = 3$)	0.868	0.822	0.028	0.003

Table 20: Performance and fairness metrics for baseline Logit and FAN models on the 2019 test data. The fairness metrics are reported as absolute differences in group-specific rates between minority ($a = 1$) and non-minority ($a = 0$).

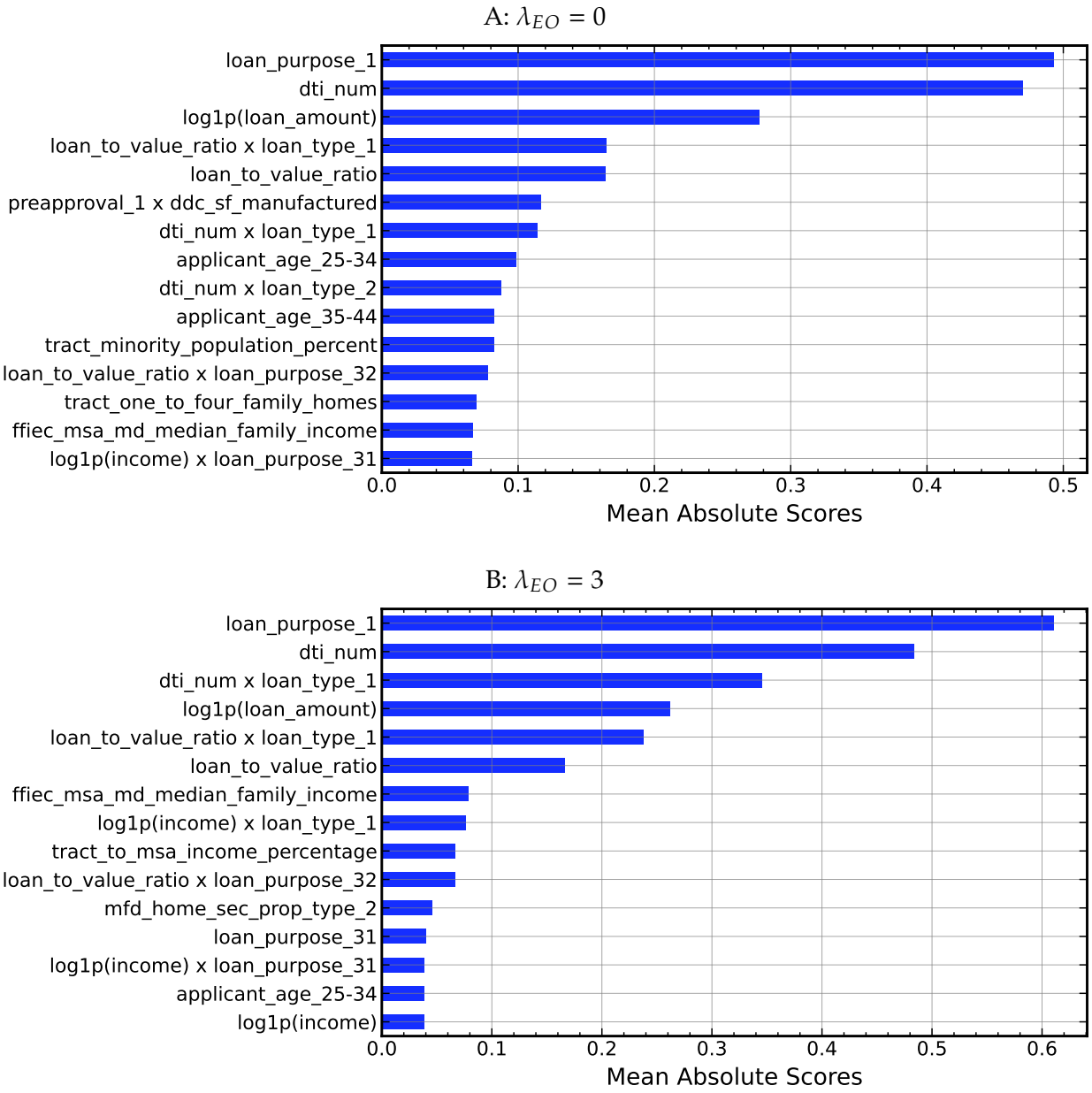


Figure 14: Mean Absolute Scores $S(i)$ and $S(i, j)$

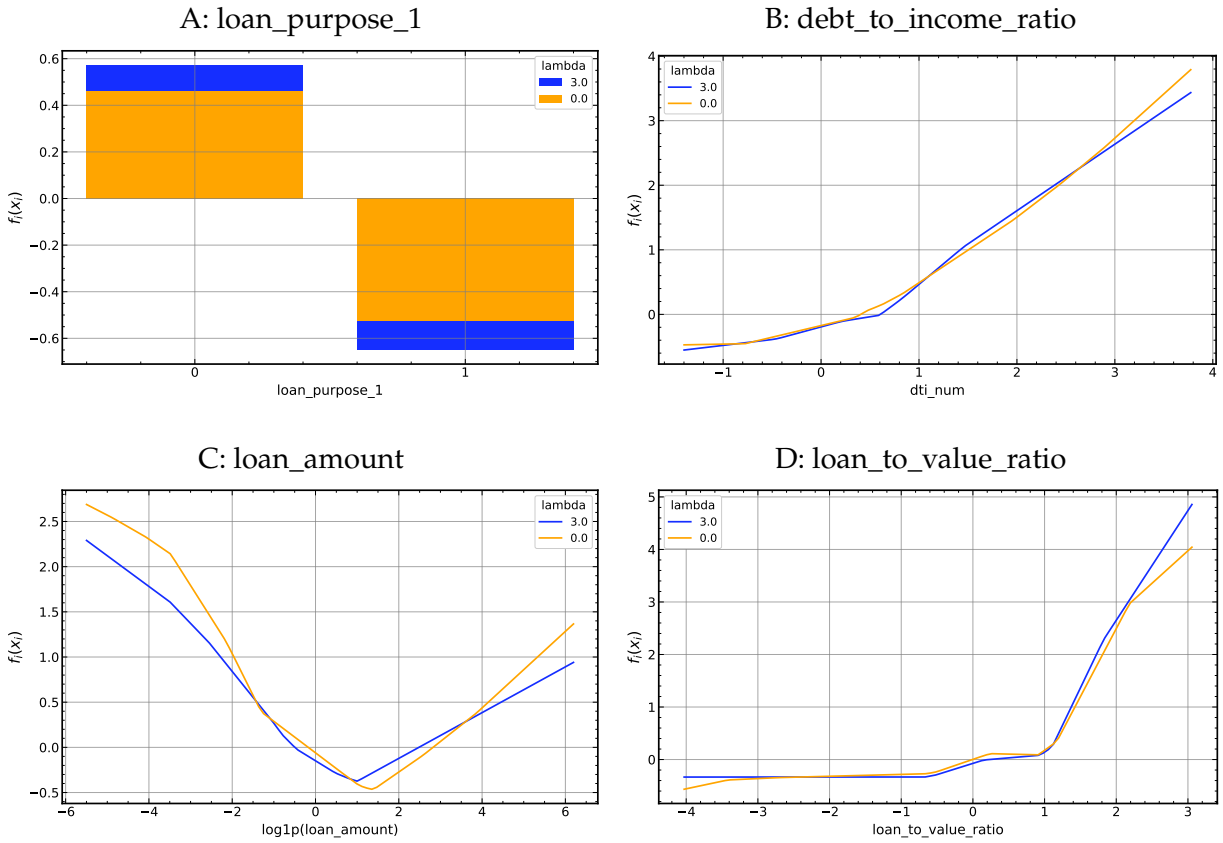


Figure 15: Univariate shape functions $f_i(x_i)$

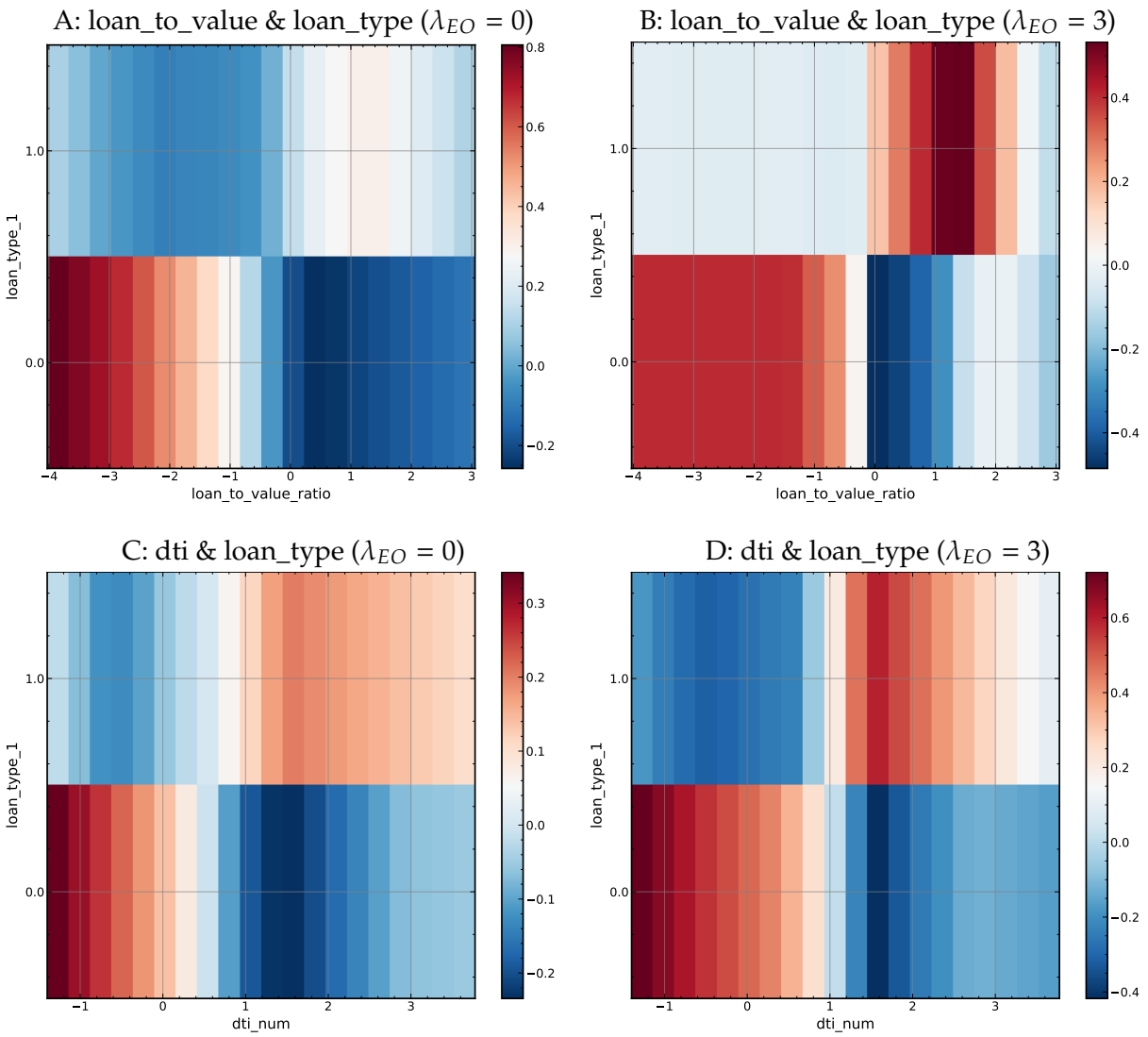


Figure 16: Interaction effects $f_{ij}(x_i, x_j)$

K Robustness: Bank-lender jumbo loan subsample

Additionally, we conduct a robustness exercise designed to address the concern that the main empirical results may partly reflect agency underwriting rules, secondary-market eligibility, or the business models of nonbank mortgage originators rather than lender-specific underwriting discretion. To move toward a setting in which approval decisions are more plausibly shaped by bank-specific underwriting, we re-estimate the models on a restricted sample of conventional nonconforming jumbo applications reported by depository bank and thrift lenders. Starting from the main HMDA application sample, we retain only first-lien, owner-occupied, site-built, one-unit, conventional home-purchase applications and then restrict to conventional nonconforming applications using `conforming_loan_limit = NC`. Finally, we merge the application data to the Philadelphia Fed HMDA Lender File by lender legal entity identifier (`LEI`) and retain only lenders whose type is in the depository bank/thrift lenders category. Because this remains an application-level sample, both denied and originated applications are preserved, so the original deny-versus-originate prediction problem is unchanged.

Table 21 shows that our main results are robust to these restrictions. Predictive performance remains strong throughout: test accuracy stays between 0.911 and 0.916, and ROC–AUC remains between 0.797 and 0.806, outperforming the baseline Logit model. At the same time, the fairness-constrained models continue to reduce disparity with only minimal predictive cost. While improvements are not strictly monotone, relative to the unconstrained baseline, the specification with fairness weight $\lambda = 3$ lowers the TPR gap from 0.0257 to 0.0220 and the FPR gap from 0.0040 to 0.0029, while reducing test accuracy by only about 0.47 percentage points and ROC–AUC by only about 0.009. In other words, the central pattern from the main analysis survives the bank-jumbo restriction: meaningful fairness improvements can still be obtained at negligible cost to predictive fit.

	Accuracy	ROC–AUC	$\lvert TPR_1 - TPR_0 \rvert$	$\lvert FPR_1 - FPR_0 \rvert$
Logit (baseline)	0.792	0.777	0.090	0.066
FAN ($\lambda_{EO} = 0$)	0.916	0.806	0.026	0.004
FAN ($\lambda_{EO} = 1$)	0.913	0.802	0.038	0.002
FAN ($\lambda_{EO} = 2$)	0.912	0.797	0.029	0.002
FAN ($\lambda_{EO} = 3$)	0.911	0.798	0.022	0.002

Table 21: Performance and fairness metrics for baseline Logit and FAN models on the subsample test data. The fairness metrics are reported as absolute differences in group-specific rates between minority ($a = 1$) and non-minority ($a = 0$).

L Robustness: Subgroup fairness

We also examine whether the overall fairness improvement translates into improvements for certain subgroups of applicants. For each model, we fix a single decision threshold τ on the test set so that the model's overall approval share matches the observed approval share (recall that y=1 denotes rejection, and predicted rejection is $\widehat{y} = \mathbb{1}\{\hat{p} \geq \tau\}$). Within each subgroup we compute $\text{TPR}_g = \Pr(\widehat{y} = 1 \mid Y = 1, A = g)$ and $\text{FPR}_g = \Pr(\widehat{y} = 1 \mid Y = 0, A = g)$ for $g \in \{0, 1\}$ (non-minority, minority), together with the gaps $\Delta\text{TPR} = \text{TPR}_1 - \text{TPR}_0$ and $\Delta\text{FPR} = \text{FPR}_1 - \text{FPR}_0$. We summarize the equalized-odds gap for a subgroup as $\max\{|\Delta\text{TPR}|, |\Delta\text{FPR}|\}$ and report results for intersections of race with sex, loan purpose, and loan type, as well as geographic subgroups formed by the top 25 MSAs by volume. Subgroups with undefined rates (for example, no positives for at least one group when computing ΔTPR) are excluded from the corresponding cell. Counts by outcome are shown in the detailed tables.

Table 22 reports the 95$^{\text{th}}$ percentile of the subgroup-level equalized-odds gap across each partition. This provides a tail-risk summary of subgroup fairness, i.e., 95% of subgroups have EO gaps no larger than the reported magnitude. Moving from λ=0 to λ=2 yields sizable improvements in the most problematic partitions. The 95$^{\text{th}}$ percentile gap for race×purpose falls from 0.520 to 0.378, and for race×loan_type from 0.199 to 0.189 with a further decline to 0.116 at λ=3. Race×sex shows small but consistent reductions (from 0.192 to about 0.186). For MSAs, the 95$^{\text{th}}$ percentile decreases from 0.179 at λ=0 to 0.162 at λ=2, with a small uptick to 0.166 at λ=3. These patterns indicate that the fairness penalty primarily mitigates disparities concentrated along product and program dimensions, with more modest changes along the sex dimension.

Tables 23 and 24 provide subgroup-level detail for λ=0 and λ=3. Within the sex partition (Table 23), the two largest subgroups exhibit clear improvements. For the first subgroup, the equalized-odds gap declines from 0.0610 to 0.0268, driven by reductions in both ΔTPR and ΔFPR. For the second subgroup, the gap falls from 0.0919 to 0.0488, with the false-positive differential shrinking from 0.0441 to 0.0177. In contrast, for a very small subgroup (N=37) the gap remains dominated by the true-positive difference (ΔTPR=0.20 in both models), even though the false-positive difference compresses substantially. This illustrates that extremely small cells can limit the effective strength of the in-sample fairness signal for one of the two components.

Turning to loan purpose (Table 24), the largest gaps at λ=0 occur in categories associated with non-purchase activity, and these decline markedly under the fairness penalty. For example, the gap for purpose code 4 falls from 0.2452 to 0.1499, and for purpose code

31 from 0.1529 to 0.1009. The gap for purpose code 32 also declines from 0.1388 to 0.0709. The home-purchase category (purpose code 1) is close to parity at baseline and improves further to 0.0268. The very small purpose code 5 cell shows a reduction but remains volatile due to the limited sample size. Overall, these results show that the fairness regularization reduces subgroup disparities where they are most pronounced, and that the reductions are achieved without flattening core risk signals, consistent with our earlier decomposition that shows a reweighting away from proxy-like variables and sharp interactions toward fundamental underwriting channels.

The MSA analysis, summarized in Table 22, shows that improvements at the geographic level are also present. The 95$^{\text{th}}$ percentile of the equalized-odds gap across the top 25 MSAs declines when moving from λ=0 to λ=2, with a slight reversal at λ=3. As with the purpose and loan-type partitions, the largest gains occur at intermediate levels of the fairness weight, which is consistent with the trade-off between shrinking cross-group error-rate gaps and preserving fine-grained sorting at the approval frontier. Across all partitions, the subgroup evidence indicates that the aggregate fairness improvement is not masking offsetting increases in intersectional slices. For exposition, Tables 23 and 24 present the detailed underlying results for sex and purpose subgroups, respectively. They illustrate how gaps shrink, particularly within the largest subgroups, when penalization is imposed.

Model	95$^{\text{th}}$ pct (race×sex)	95$^{\text{th}}$ pct (race×purpose)	95$^{\text{th}}$ pct (race×loan_type)	95$^{\text{th}}$ pct (top-25 MSAs)
FAN (λ = 0)	0.1924	0.5196	0.1988	0.1786
FAN (λ = 1)	0.1878	0.5133	0.1964	0.1759
FAN (λ = 2)	0.1853	0.3777	0.1887	0.1618
FAN (λ = 3)	0.1866	0.3708	0.1164	0.1656

Table 22: EO gaps within subgroups. This table presents observed EO disparities within subgroups, reporting the 95th percentile within each cell.

Subgroup	N	$N_{Y=1}$	$N_{Y=0}$	TPR_0	FPR_0	TPR_1	FPR_1	ΔTPR	$\lvert\Delta\text{TPR}\rvert$	ΔFPR	$\lvert\Delta\text{FPR}\rvert$	EO gap (max Δ)
Panel A: $\lambda_{EO} = 0$												
sex=6	37	10	27	0.6000	0.1053	0.8000	0.2500	0.2000	0.2000	0.1447	0.1447	0.2000
sex=3	221	70	151	0.7727	0.0932	0.8077	0.2424	0.0350	0.0350	0.1492	0.1492	0.1492
sex=2	25433	6349	19084	0.5987	0.1202	0.6907	0.1643	0.0919	0.0919	0.0441	0.0441	0.0919
sex=1	49309	10698	38611	0.5499	0.1092	0.6109	0.1497	0.0610	0.0610	0.0405	0.0405	0.0610
Panel B: $\lambda_{EO} = 3$												
sex=6	37	10	27	0.6000	0.1053	0.8000	0.1250	0.2000	0.2000	0.0197	0.0197	0.2000
sex=3	221	70	151	0.7727	0.1017	0.8077	0.2121	0.0350	0.0350	0.1104	0.1104	0.1104
sex=2	25433	6349	19084	0.6115	0.1213	0.6603	0.1390	0.0488	0.0488	0.0177	0.0177	0.0488
sex=1	49309	10698	38611	0.5627	0.1158	0.5895	0.1298	0.0268	0.0268	0.0140	0.0140	0.0268

Table 23: EO gaps within sex subgroups

Subgroup	N	$N_{Y=1}$	$N_{Y=0}$	TPR_0	FPR_0	TPR_1	FPR_1	ΔTPR	$\lvert\Delta\text{TPR}\rvert$	ΔFPR	$\lvert\Delta\text{FPR}\rvert$	EO gap (max Δ)
Panel A: $\lambda_{EO} = 0$												
purpose=5	28	13	15	0.3333	0.2222	0.5000	0.8333	0.1667	0.1667	0.6111	0.6111	0.6111
purpose=4	2176	1137	1039	0.8233	0.4796	0.9097	0.7248	0.0864	0.0864	0.2452	0.2452	0.2452
purpose=2	2122	1079	1043	0.8345	0.5372	0.9591	0.7261	0.1246	0.1246	0.1889	0.1889	0.1889
purpose=31	11218	3630	7588	0.5728	0.2417	0.6758	0.3946	0.1030	0.1030	0.1529	0.1529	0.1529
purpose=32	17365	5433	11932	0.5771	0.2072	0.6922	0.3460	0.1151	0.1151	0.1388	0.1388	0.1388
purpose=1	42091	5835	36256	0.4517	0.0267	0.4947	0.0408	0.0430	0.0430	0.0142	0.0142	0.0430
Panel B: $\lambda_{EO} = 3$												
purpose=5	28	13	15	0.3333	0.2222	0.5000	0.6667	0.1667	0.1667	0.4444	0.4444	0.4444
purpose=4	2176	1137	1039	0.8422	0.5290	0.8958	0.6789	0.0537	0.0537	0.1499	0.1499	0.1499
purpose=2	2122	1079	1043	0.8507	0.5688	0.9327	0.7070	0.0820	0.0820	0.1382	0.1382	0.1382
purpose=31	11218	3630	7588	0.5851	0.2457	0.6316	0.3466	0.0465	0.0465	0.1009	0.1009	0.1009
purpose=32	17365	5433	11932	0.6017	0.2277	0.6676	0.2985	0.0658	0.0658	0.0709	0.0709	0.0709
purpose=1	42091	5835	36256	0.4510	0.0241	0.4778	0.0308	0.0268	0.0268	0.0067	0.0067	0.0268

Table 24: EO gaps within loan purpose subgroups

M Robustness: Decomposition of the global score gap

In Section 5.2, we examined the drivers of the EO disparity in the neighborhood of the $s_\theta(X) = 0$ region, where the score decomposition explains output probabilities. For robustness, we also investigate which features generate the EO disparity in our models and how λ_{EO} changes that composition on the full test data. The goal is to attribute cross-group differences in the model's decision boundary to individual features of HMDA loan applications in a way that is exact for our additive architecture and comparable across fairness weights.

As defined above, the predicted denial probability is $\hat{p}_\theta(x) = \sigma\big(s_\theta(x)\big)$, where $\sigma(\cdot)$ the logistic link function, but we explain predictions on the score ($s_\theta(x)$) scale, where additivity holds exactly and each term's contribution is well-defined. The (soft) EO target equalizes across groups the class-conditional average probabilities $\mu_{g,y}(\theta) := \mathbb{E}\left[\sigma\big(s_\theta(X)\big) \,\middle|\, A{=}g,\ Y{=}y\right]$. We now decompose the *class-conditional mean score gap* between groups into per-term contributions. For any main or interaction term $t \in \{i\} \cup \{(i,j)\}$, we define

$$\bar{f}_t^{\,a,y} \;:=\; \mathbb{E}\big[f_t(X) \,\big|\, A{=}a,\ Y{=}y\big]\,, \qquad \Delta_t^{(y)} \;:=\; \bar{f}_t^{\,1,y} \;-\; \bar{f}_t^{\,0,y}.$$

Summing over all additive terms yields the group difference in the mean score within label y,

$$\mathbb{E}\big[s_\theta(X) \mid A{=}1,\ Y{=}y\big] - \mathbb{E}\big[s_\theta(X) \mid A{=}0,\ Y{=}y\big] \;=\; \sum_t \Delta_t^{(y)}.$$

The vector $\{\Delta_t^{(y)}\}_t$ is therefore an exact attribution of the class-conditional score gap into economically interpretable term-wise components, without post-hoc approximation error. For each fitted model (indexed by the fairness weight λ_{EO}) and each term t, we compute application-level contributions $f_t(X)$ on the test set and then average them conditional on (A, Y) to obtain $\bar{f}_t^{\,A,y}$.

Large positive $\Delta_t^{(1)}$ (within $Y{=}1$ denials) indicate that, among denied applicants, minorities load more heavily on term t than otherwise similar non-minorities (e.g., a sizable positive DTI delta means denied minority applicants are assigned higher DTI-induced scores on average). Conversely, large negative values imply the term favors minorities within that outcome group. The same interpretation applies to $Y{=}0$ originations. Comparing the decomposition across λ_{EO} reveals how fairness training operates: If large Δ_t^y shrink, the model is aligning class-conditional means for these prediction terms.

Figure 17 reports the ten largest $\Delta_t^{(y)}$ (by absolute value) for each model and observed class label. In the form of `dti`, repayment capacity is the dominant channel among denials

and remains intact under fairness. For Y=1 (denied), the largest signed contribution at λ=0 is `dti_num` with $\Delta^{(1)} = 0.158$, indicating that, among denied applicants, minorities carry a higher DTI-induced score on average than non-minorities. Under λ=3 this term remains the largest and only slightly decreases ($\Delta^{(1)} = 0.153$), consistent with the fairness penalty preserving the core underwriting signal while aligning other channels. A similar pattern can be observed for Y=0 (originated): `dti_num` remains positive and of comparable size (from 0.090 to 0.087), suggesting that repayment capacity continues to influence the score gap after fairness is imposed.

In contrast, geographic proxies seem to compress sharply. At λ=0, `tract_minority_population_percent` is among the largest contributors in both outcome groups ($\Delta^{(1)}$ = 0.141 among denials and $\Delta^{(0)}$ = 0.127 among originations), pointing to a meaningful role for neighborhood composition in the raw model. With λ=3, these loads fall substantially (to 0.039 for Y=1 and 0.036 for Y=0), indicating that the fairness constraint reduces reliance on geographic correlates, which might plausibly proxy for minority membership, while leaving the repayment-capacity and collateral channels in place. Other location-income controls (`ffiec_msa_md_median_family_income`, `tract_to_msa_income_percentage`) remain small and comparatively stable (roughly 0.01-0.03 across outcomes and λ), consistent with a broad reduction of proxy effects rather than a simple reallocation to other geographic variables.

We also observe that collateral and size terms contribute moderately and with consistent signs. The main `loan_to_value_ratio` term is positive in both groups and becomes slightly larger at λ=3 ($\Delta^{(1)}$: 0.033 to 0.036; $\Delta^{(0)}$: 0.035 to 0.040), indicating that differences in LTV placement continue to explain part of the EO gap after fairness is imposed. Balance and collateral jointly matter through `log1p(loan_amount)` and `log1p(property_value)`, which carry small positive loads in both groups. Their interaction `log1p(loan_amount)`×`log1p(property_value)` remains among the top terms for originations ($\Delta^{(0)}$: 0.035 at λ=0, 0.032 at λ=3), pointing to an exposure-collateral channel consistent with the interaction shape function.

Additionally, programmatic and product channels are informative. `loan_purpose_1` (purchase) enters with a negative sign in both outcomes (about −0.065 for Y=1 and −0.065 to −0.068 for Y=0), implying that, conditional on approval status, the purchase purpose reduces the score gap for minorities relative to non-minorities. The interaction `dti_num` × `loan_type_1` (conventional) is negative and modest in both groups (about −0.013 for Y=1 and −0.022 to −0.019 for Y=0), consistent with relatively less adverse DTI impact in insured/guaranteed channels. Process- and build-type interactions appear only at small magnitudes: for Y=0 they include `preapproval_1` × `construction_method_2` at

λ=0 (−0.021) and `preapproval_2` × `manufactured_home_secured_property_type_2` at λ=3 (−0.015), indicating that EO alignment is not achieved via a single dominating interaction. Finally, age composition effects are minor, with `applicant_age_25-34` showing small negative loads (about −0.007) in denials at both λ.

Overall, across both observed outcomes, the EO penalty meaningfully rebalances the composition of the conditional score gap: it substantially reduces reliance on a strong geographic correlate while preserving the central, economically grounded channels such as repayment capacity and collateral/size at moderate magnitudes. The residual gaps are dispersed across core underwriting determinants (DTI, LTV, purpose), rather than concentrated in idiosyncratic process interactions, indicating improved parity without flattening fundamental risk signals.

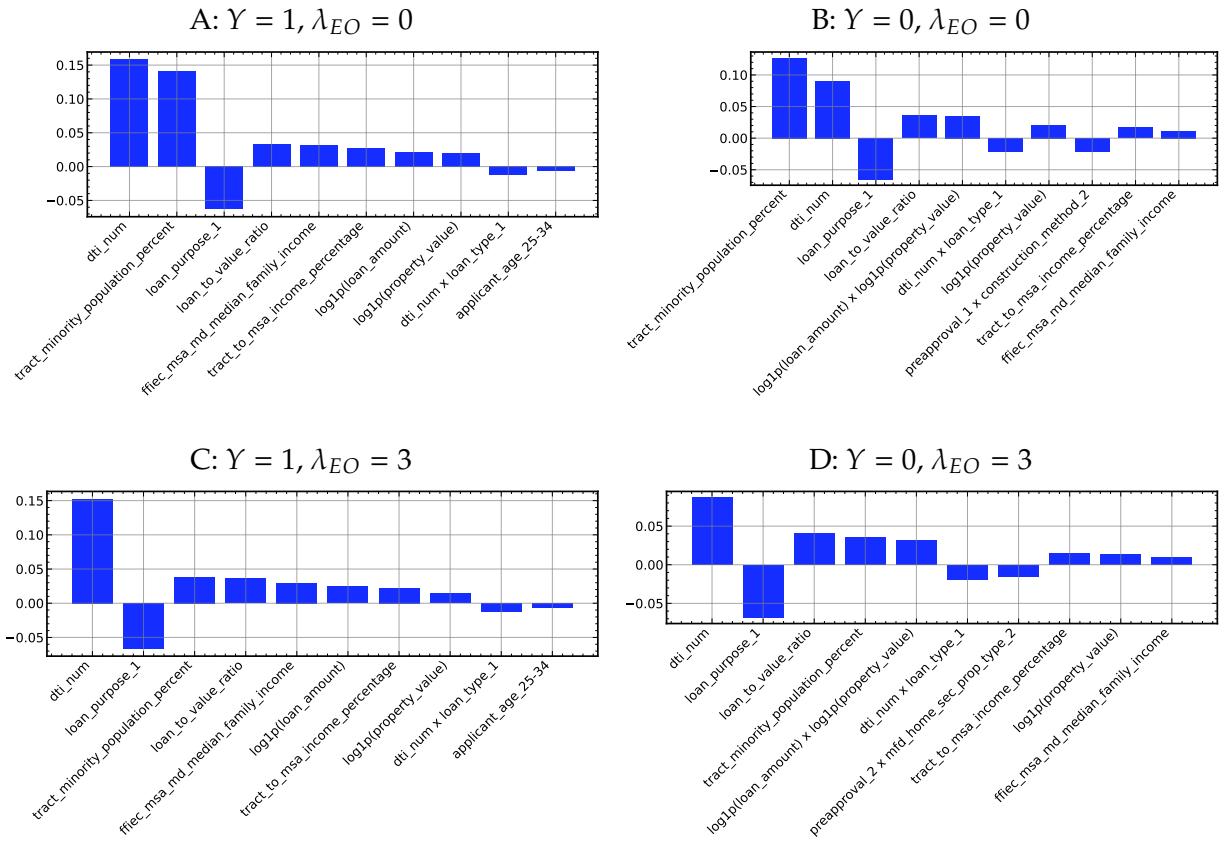


Figure 17: $\Delta_t^{(y)}$ across class labels and models. This figure compares the most important contributors to the score gap for different outcomes and penalization levels. For each case, the ten largest $\Delta_t^{(y)}$ (by absolute value) are reported.

References

Adelino, Manuel, Antoinette Schoar, and Felipe Severino, 2025, Credit supply and house prices: Evidence from mortgage market segmentation, *Journal of Financial Economics* 163, 103958.

Agarwal, Sumit, Itzhak Ben-David, and Vincent Yao, 2017, Systematic mistakes in the mortgage market and lack of financial sophistication, *Journal of Financial Economics* 123, 42–58.

Alvarez-Melis, David, and Tommi S. Jaakkola, 2018, Towards robust interpretability with self-explaining neural networks, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 7786–7795 (Curran Associates Inc., Red Hook, NY, USA).

Ambrose, Brent W., James N. Conklin, and Luis A. Lopez, 2021, Does borrower and broker race affect the cost of mortgage credit?, *Review of Financial Studies* 34, 790–826.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace, 2022a, Algorithmic discrimination and input accountability under the Civil Rights Acts, *Berkeley Technology Law Journal* 36, 675–736.

Bartlett, Robert, Adair Morse, Richard Stanton, and Nancy Wallace, 2022b, Consumer-lending discrimination in the FinTech Era, *Journal of Financial Economics* 143, 30–56.

Bayer, Patrick, Fernando Ferreira, and Stephen L. Ross, 2018, What drives racial and ethnic differences in high-cost mortgages? The role of high-risk lenders, *Review of Financial Studies* 31, 175–205.

Bell, Sebastian, Ali Kakhbod, Martin Lettau, and Abdolreza Nazemi, 2025, Glass box machine learning and corporate bond returns, Working Paper 33320, National Bureau of Economic Research.

Berg, Tobias, Valentin Burg, Ana Gombović, and Manju Puri, 2020, On the rise of FinTechs: Credit scoring using digital footprints, *Review of Financial Studies* 33, 2845–2897.

Bhutta, Neil, and Aurel Hizmo, 2021, Do minorities pay more for mortgages?, *Review of Financial Studies* 34, 763–789.

Black, Emily, John Logan Koepke, Pauline T. Kim, Solon Barocas, and Mingwei Hsu, 2024, Less discriminatory algorithms, *Georgetown Law Journal* 113, 53–120.

Black, Harold, Robert L. Schweitzer, and Lewis Mandell, 1978, Discrimination in mortgage lending, *American Economic Review* 68, 186–191.

Black, Harold A., Thomas P. Boehm, and Ramon P. DeGennaro, 2003, Is there discrimination in mortgage pricing? The case of overages, *Journal of Banking and Finance* 27, 1139–1165.

Blinder, Alan S., 1973, Wage discrimination: Reduced form and structural estimates, *Journal of Human Resources* 8, 436–455.

Board of Governors of the Federal Reserve System, 2011, Supervisory guidance on model risk management, Technical Report SR 11-7, Federal Reserve Board.

Brainard, Lael, 2021, Supporting responsible use of AI and equitable outcomes in financial services.

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu, 2025, Forest through the trees: Building cross-sections of stock returns, *Journal of Finance* 80, 2447–2506.

Butler, Alexander W., Erik J. Mayer, and James P. Weston, 2023, Racial disparities in the auto loan market, *Review of Financial Studies* 36, 1–41.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2024a, Deep learning in asset pricing, *Management Science* 70, 714–750.

Chen, Yifei, Bryan T. Kelly, and Dacheng Xiu, 2024b, Expected returns and large language models, Working paper, University of Chicago.

Cheng, Ping, Zhenguo Lin, and Yingchun Liu, 2015, Racial discrepancy in mortgage interest rates, *Journal of Real Estate Finance and Economics* 51, 101–120.

Consumer Financial Protection Bureau, 2022, Consumer Financial Protection Circular 2022-03: Adverse action notification requirements in connection with credit decisions based on complex algorithms.

Courchane, Marsha, and David Nickerson, 1997, Discrimination resulting from overage practices, *Journal of Financial Services Research* 11, 133–151.

Das, Sanjiv R., Richard Stanton, and Nancy Wallace, 2023, Algorithmic fairness, *Annual Review of Financial Economics* 15, 565–593.

DeFusco, Anthony A., Stephanie Johnson, and John Mondragon, 2020, Regulating household leverage, *Review of Economic Studies* 87, 914–958.

Dobbie, Will, Andres Liberman, Daniel Paravisini, and Vikram Pathania, 2021, Measuring bias in consumer lending, *Review of Economic Studies* 88, 2799–2832.

Fallahgoul, Hasan, Vincentius Franstianto, and Xin Lin, 2024, Asset pricing with neural networks: Significance tests, *Journal of Econometrics* 238, 105574.

Folland, Gerald B., 1999, *Real Analysis: Modern Techniques and Their Applications*, second edition (Wiley, New York).

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, 2022, Predictably unequal? The effects of machine learning on credit markets, *Journal of Finance* 77, 5–47.

Fuster, Andreas, Stephanie H. Lo, and Paul S. Willen, 2024, The time-varying price of financial intermediation in the mortgage market, *The Journal of Finance* 79, 2553–2602.

Gerardi, Kristopher, Paul S. Willen, and David Hao Zhang, 2023, Mortgage prepayment, race, and monetary policy, *Journal of Financial Economics* 147, 498–524.

Ghent, Andra C., Rubén Hernández-Murillo, and Michael T. Owyang, 2014, Differences in subprime loan pricing across races and neighborhoods, *Regional Science and Urban Economics* 48, 199–215.

Gillis, Talia B., and Jann L. Spiess, 2019, Big data and discrimination, *University of Chicago Law Review* 86, 459–487.

Green, Richard K., and Susan M. Wachter, 2005, The American mortgage in historical and international context, *Journal of Economic Perspectives* 19, 93–114.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2021, Autoencoder asset pricing models, *Journal of Econometrics* 222, 429–450.

Gurun, Umit, Gregor Matvos, and Amit Seru, 2016, Advertising expensive mortgages, *Journal of Finance* 71, 2371–2416.

Hanson, Andrew, Zackary Hawley, Hal Martin, and Bo Liu, 2016, Discrimination in mortgage lending: Evidence from a correspondence experiment, *Journal of Urban Economics* 92, 48–65.

Hardt, Moritz, Eric Price, and Nati Srebro, 2016, Equality of opportunity in supervised learning, in D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, volume 29, 3315–3323.

Hastie, Trevor, and Robert Tibshirani, 1986, Generalized additive models, *Statistical Science* 1, 297–310.

Hellman, Deborah, 2020, Measuring algorithmic fairness, *Virginia Law Review* 106, 811–866.

Howell, Sabrina T., Theresa Kuchler, David Snitkoff, Johannes Stroebel, and Jun Wong, 2024, Lender automation and racial disparities in credit access, *Journal of Finance* 79, 1457–1512.

Hurlin, Christophe, Christophe Pérignon, and Sébastien Saurin, 2025, The fairness of credit scoring models, *Management Science* (forthcoming).

Jansen, Mark, Hieu Quang Nguyen, and Amin Shams, 2025, Rise of the machines: The impact of automated underwriting, *Management Science* 71, 955–975.

Kakhbod, Ali, Leonid Kogan, Peiyao Li, and Dimitris Papanikolaou, 2025, Measuring creative destruction, Working paper, University of California, Berkeley.

Keys, Benjamin J., Devin G. Pope, and Jaren C. Pope, 2016, Failure to refinance, *Journal of Financial Economics* 122, 482–499.

Kim, Pauline, 2017, Auditing algorithms for discrimination, *University of Pennsylvania Law Review Online* 166, 189–203.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein, 2020, Algorithms as discrimination detectors, *Proceedings of the National Academy of Sciences* 117, 30096–30100.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan, 2017, Inherent trade-offs in the fair determination of risk scores, in Christos H. Papadimitriou, ed., *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 431–432 (Schloss Dagstuhl - Leibniz-Zentrum für Informatik).

Kroll, Joshua, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson, and Harlan Yu, 2017, Accountable algorithms, *University of Pennsylvania Law Review* 165, 633–705.

Lee, Michelle Seng Ah, and Luciano Floridi, 2021, Algorithmic fairness in mortgage lending: From absolute conditions to relational trade-offs, *Minds and Machines* 31, 165–191.

Liu, Feng, Jason Dietrich, Young Jo, and Misha Davies, 2019, Introducing new and revised data points in HMDA, Working Paper 19-5, Consumer Financial Protection Bureau, Rochester, NY.

Lopez-Lira, Alejandro, and Nikolai L. Roussanov, 2023, Do common factors really explain the cross-section of stock returns?, Working paper, University of Pennsylvania.

Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker, 2013, Accurate intelligible models with pairwise interactions, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, 623–631 (Association for Computing Machinery, New York, NY, USA).

McDiarmid, Colin, 1989, On the method of bounded differences, in J. Siemons, ed., *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, London Mathematical Society Lecture Note Series, 148–188 (Cambridge University Press).

Munnell, Alicia H., Lynn Browne, James McEneaney, and Geoffrey Tootel, 1996, Mortgage lending in Boston: Interpreting HMDA data, *American Economic Review* 86, 25–54.

Oaxaca, Ronald, 1973, Male-female wage differentials in urban labor markets, *International Economic Review* 14, 693–709.

O'Neil, Cathy, 2016, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, New York).

Reid, Carolina K., Debbie Bocian, Wei Li, and Roberto G. Quercia, 2017, Revisiting the subprime crisis: The dual mortgage market and mortgage defaults by race and ethnicity, *Journal of Urban Affairs* 39, 469–487.

Rudin, Cynthia, 2019, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* 1, 206–215.

Stasinopoulos, Mikis D., Thomas Kneib, Nadja Klein, Andreas Mayr, and Gillian Z. Heller, 2024, *Generalized Additive Models for Location, Scale and Shape: A Distributional Regression Approach, with Applications*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge).

Sudjianto, Agus, and Aijun Zhang, 2021, Designing inherently interpretable machine learning models, in *ACM ICAIF 2021 Workshop on Explainable AI in Finance, November 3, 2021*.

Treadway v. Gateway Chevrolet Oldsmobile Inc., 2004, 362 F.3d 971 (7th Cir. 2004).

Ustun, Berk, and Cynthia Rudin, 2019, Learning optimized risk scores, *Journal of Machine Learning Research* 20, 1–75.

van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira, 2023, Man versus machine learning: The term structure of earnings expectations and conditional biases, *Review of Financial Studies* 36, 2361–2396.

Willen, Paul, and David Hao Zhang, 2025, Do lenders still discriminate? A robust approach for assessing differences in menus, Working Paper 20-19, Federal Reserve Bank of Boston.

Yang, Crystal S., and Will Dobbie, 2020, Equal protection under algorithms: A new statistical and legal framework, *Michigan Law Review* 119, 291–396.

Yang, Zebin, Aijun Zhang, and Agus Sudjianto, 2021, GAMI-Net: An explainable neural network based on generalized additive models with structured interactions, *Pattern Recognition* 120, 108192.