FAIRNESS BY DESIGN: MACHINE LEARNING AND INTERPRETABLE MORTGAGE LENDING

Anonymous authors
Paper under double-blind review

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.

1 INTRODUCTION

Machine learning (ML) can improve prediction in consumer lending, but credit decisions face unusually strong requirements for fairness and transparency. U.S. fair-lending and model-risk regimes—including the Equal Credit Opportunity Act (ECOA), Regulation B, the Fair Credit Reporting Act (FCRA), and SR 11-7—require that credit decisions be well documented and that applicants receive specific reasons for adverse action (Board of Governors of the Federal Reserve System, 2011; Consumer Financial Protection Bureau, 2022). In particular, Regulation B requires that these reasons accurately reflect the factors actually considered by the creditor, and recent CFPB guidance emphasizes that this obligation continues to apply when lenders use complex algorithms (Consumer Financial Protection Bureau, 2022). These requirements create a tension between predictive flexibility and regulatory explainability: simple linear scorecards are easy to inspect, whereas flexible "black-box" models may be difficult to explain at the level of an individual decision (Ustun & Rudin, 2019; Rudin, 2019).

Transparency alone, however, does not guarantee equitable decisions. Protected characteristics can enter predictions indirectly through correlated proxies (Gillis & Spiess, 2019), and greater model flexibility can alter the distribution of credit outcomes across groups. In mortgage markets, Bartlett et al. (2022) document higher prices for risk-equivalent Black and Hispanic borrowers, while Fuster et al. (2022) show that flexible ML can increase dispersion in predicted risk and worsen some within-group outcomes even when average access improves. These findings motivate a model-design problem that is particularly important in regulated lending: can a decision rule be nonlinear enough to capture economically meaningful structure, explicitly constrained to reduce group disparities, and still transparent enough to explain individual decisions?

We address this question by developing the *Fairness-aware Additive Network* (FAN), an interpretable classifier that incorporates equalized-odds (EO) fairness directly into estimation. FAN combines a differentiable EO penalty with an additive, globally separable architecture. The fairness constraint therefore changes the learned decision rule itself rather than operating through ex-post reweighting or threshold adjustments, while the additive structure decomposes every prediction exactly into feature-level contributions. In the spirit of inherently interpretable and self-explaining models (Alvarez-Melis & Jaakkola, 2018; Rudin, 2019; Sudjianto & Zhang, 2021), FAN provides global shape functions and faithful instance-level attributions that can be mapped into adverse-action-style explanations without relying on post-hoc explanation methods.

We interpret FAN as a candidate less discriminatory alternative (LDA) within a common public-information set. The objective is to identify a decision rule that simultaneously reduces minority–

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

non-minority disparities in errors relative to observed underwriting outcomes, preserves strong out-of-sample fit to those outcomes, and remains transparent by construction. This framing is especially relevant when multiple predictive models can fit the same data but generate different distributions of decisions across groups (Black et al., 2024).

Our contribution is threefold. First, we introduce an in-processing EO regularizer with a direct connection to decision-relevant disparities. Although the training penalty is smooth and does not require fixing a classification threshold, we show that it places greatest weight near the decision boundary, where small score changes alter approval decisions. We further relate the penalty to threshold-indexed EO disparities and establish an out-of-sample generalization guarantee under the architectural restrictions used in FAN. These results give the fairness objective an interpretable connection to operational lending decisions rather than treating it as a purely statistical penalty.

Second, FAN shows that fairness constraints and inherent interpretability can be combined without reverting to a linear scorecard. Neural shape functions capture nonlinear effects and a sparse set of interactions, while additivity preserves an exact decomposition of every prediction. The model therefore makes it possible to inspect both the underwriting rule itself and how that rule changes when fairness is incorporated into estimation.

Third, we apply FAN to public Home Mortgage Disclosure Act (HMDA) approval and denial decisions and find a favorable fairness–performance trade-off. Increasing the EO penalty materially reduces minority–non-minority gaps in model-implied error rates with negligible loss in predictive performance. The model remains economically interpretable: fairness regularization reduces reliance on some geographic variables correlated with protected status while preserving the importance of core underwriting signals such as debt-to-income (DTI) and loan-to-value (LTV). Around institutional underwriting cutoffs, minority-specific shortfalls decline as the fairness penalty increases, and the improvements also appear within lenders rather than being driven primarily by reallocation across institutions. At the same time, we find little evidence that the fairness adjustment disrupts conditional pricing relationships among approved borrowers.

Taken together, the results show that compliance-relevant explainability need not require purely linear models and that greater algorithmic flexibility need not come at the cost of more disparate decisions. Treating fairness and interpretability as design objectives can produce less disparate mortgage decision rules while preserving predictive performance and a transparent connection between model inputs and individual outcomes.

2 Related Work

Our analysis connects work on machine learning in finance, mortgage-market disparities, and algorithmic fairness. Flexible ML methods can improve financial prediction (Gu et al., 2021; Bell et al., 2025), but in credit markets they can also change the distribution of predicted risk and outcomes across borrowers (Fuster et al., 2022). Related work on automated credit scoring emphasizes both predictive gains and the possibility that high-dimensional models reproduce disparities through variables correlated with protected status (Berg et al., 2020; Howell et al., 2024). A parallel mortgage literature documents persistent racial and ethnic differences in pricing and access to credit (Bayer et al., 2018; Bartlett et al., 2022). We complement this evidence by asking whether the underwriting rule itself can be redesigned to reduce disparities while retaining nonlinear predictive structure and exact interpretability.

The paper also relates to work on interpretable machine learning and algorithmic fairness. Inherently interpretable and self-explaining models seek to make the learned decision rule itself understandable rather than explaining a black box after estimation (Alvarez-Melis & Jaakkola, 2018; Rudin, 2019; Sudjianto & Zhang, 2021). Equalized odds provides a formal criterion for balancing group-conditional classification errors (Hardt et al., 2016), while impossibility results show that different fairness criteria generally cannot be satisfied simultaneously (Kleinberg et al., 2017). Legal and economic work further emphasizes that excluding protected characteristics does not eliminate discrimination when correlated proxies remain (Gillis & Spiess, 2019), and that model multiplicity can create feasible less discriminatory alternatives (Black et al., 2024). Our approach connects these strands by embedding a differentiable EO objective directly in estimation while retaining a model whose predictions can be decomposed exactly. Finally, our recovered nonlinear effects and inter-

actions connect to work on institutional frictions in mortgage underwriting, where program design, guarantees, repayment capacity, and collateral affect how applications are screened (Gerardi et al., 2023). FAN allows these economically meaningful channels to remain visible while showing how their use changes under fairness regularization.

3 FAIRNESS-AWARE ADDITIVE MODELING

3.1 ARCHITECTURE

Let X_l denote the features of mortgage application l and let $y_l \in \{0, 1\}$ indicate denial (1) or origination (0). FAN predicts a denial score through the additive decomposition

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

with predicted denial probability $\hat{p}_l = \sigma(s_\theta(X_l))$. The f_i are univariate shape functions and the f_{ij} are a sparse set of pairwise interactions. Each term is parameterized by a small neural subnetwork, allowing nonlinear effects while preserving an exactly additive representation. Each function is also multiplied by a learnable scale weight (denoted w_i and v_{ij} for main and interaction effects, respectively), and those weights are regularized to encourage sparsity.

This structure provides both flexibility and inherent interpretability. Every prediction decomposes exactly into an intercept and term-level contributions, yielding global shape functions as well as faithful instance-level attributions. In lending applications, these contributions can be ranked to identify the factors pushing an application toward denial, providing a direct basis for adverse-action-style explanations rather than requiring a post-hoc explanation model.

3.2 FAIRNESS OBJECTIVE

We incorporate fairness directly into estimation using a smooth equalized-odds (EO) penalty. Let $A \in \{0, 1\}$ denote protected-group membership, with $A = 1$ for the minority group and $A = 0$ otherwise. Define soft group-conditional true- and false-positive rates by averaging predicted denial probabilities within the corresponding observed outcome classes. For two groups,

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

Training minimizes

$$\mathcal{L}(\theta) = \mathrm{BCE}(\theta) + \lambda_{\mathrm{EO}}\mathcal{P}_{\mathrm{EO}}(\theta) + \lambda_{\mathrm{main}} \sum_i |w_i| + \lambda_{\mathrm{inter}} \sum_{(i,j)\in\mathcal{I}} |v_{ij}|, \tag{3}$$

where λ_{EO} controls the fairness–fit trade-off and the additional ℓ_1-style penalties encourage sparsity in main and interaction effects.

Equalized odds is useful in this setting because mortgage underwriting is a binary decision problem in which disparities in both false rejections and false approvals are economically relevant. Importantly, although $\mathcal{P}_{\mathrm{EO}}$ is differentiable and does not require fixing an operating threshold during training, it remains closely connected to threshold-based decisions. We show that the soft group-conditional rates can be represented as weighted averages of threshold-specific rates, with the largest weight near the decision boundary. We further show that the training penalty is bounded by a threshold-averaged EO disparity measure and establish an out-of-sample generalization guarantee under the sparsity and smoothness conditions imposed on FAN. Thus, the regularizer controls a decision-relevant notion of cross-group disparity rather than serving only as an in-sample statistical criterion. Formal statements and proofs are provided in Appendix G.

3.3 ESTIMATION

FAN is estimated in three stages. We first learn the univariate main effects, then introduce a sparse set of promising pairwise interactions, and finally jointly fine-tune the complete model. This procedure preserves a compact representation while allowing economically important nonlinearities and interactions to enter the decision rule. Detailed architecture, regularization, interaction-selection, and training specifications are reported in Appendix D.

4 Data and Empirical Design

We use the 2018 release of the Home Mortgage Disclosure Act (HMDA) data, which covers more than 15 million mortgage applications and contains loan characteristics, borrower characteristics, neighborhood information, and underwriting outcomes. Following standard filters in the mortgage literature (Liu et al., 2019; Fuster et al., 2024), we restrict the sample to first-lien, single-family, owner-occupied applications. The resulting analysis sample is a representative random subsample of 500,000 applications.

The target is whether an application is denied or originated. After filtering, denials constitute about 23% of the sample and originations about 77%. This is an underwriting-outcome task rather than an ex post default-prediction task, which is important for interpretation: the model is designed to recover and improve a lending decision rule, not to estimate realized loan performance after origination.

Predictors are limited to variables plausibly available at decision time. They include loan purpose, loan amount, property value, debt-to-income ratio (DTI), loan-to-value ratio (LTV), loan term, loan type, preapproval indicators, applicant age bins, and tract-level contextual variables such as neighborhood income and minority share. Continuous variables with heavy skew, such as loan amount, property value, and income, are log transformed. Missing continuous variables are median-imputed using the training set, missing categorical variables are imputed with the modal category, and categorical features are one-hot encoded. LTV is winsorized at the 99.9th training percentile to limit the influence of implausible outliers.

For fairness evaluation we construct a binary minority indicator equal to one for Black, Hispanic, Native American, and Pacific Islander applicants and zero otherwise. The protected attribute is used for training and evaluation of the fairness term, but not as an ordinary predictive input. We split the data into 70%/15%/15% training, validation, and test sets. We compare a baseline logistic regression to FAN models estimated at $\lambda_{\mathrm{EO}} \in \{0, 1, 2, 3\}$. Hyperparameters are selected on the validation set using ROC–AUC.

The held-out test set contains 75,000 observations. Overall denials constitute 22.8% of the test sample, but the denial rate is 30.0% for minority applicants (22.1% of the sample) versus 20.8% for non-minority applicants. This class imbalance and the substantial minority–nonminority gap in denial rates make it especially important to evaluate models with group-conditional error rates and institutionally relevant boundary tests rather than relying on aggregate classification performance alone.

5 Model Performance

As the results reported in Table 1 show, the machine learning model outperforms the logistic regression for all levels of penalization, improving on both accuracy and ROC–AUC. At the same time, the machine learning approach allows us to reduce group disparities using increasing levels of penalization without reducing accuracy or ROC–AUC and while consistently beating the logit benchmark. The unconstrained FAN reaches 0.838 accuracy and 0.823 ROC–AUC. Raising the fairness weight to $\lambda_{\mathrm{EO}} = 3$ leaves those metrics essentially unchanged (0.839 and 0.823) while reducing the absolute TPR gap from 0.059 to 0.043 and the absolute FPR gap from 0.006 to 0.003. Relative to logit, the fairness-aware FAN is simultaneously more accurate and more balanced across groups.

Table 1: Out-of-sample predictive performance and fairness on the 2018 test set. Lower gap values indicate smaller minority–nonminority disparities.

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

The aggregate evidence therefore points to a relatively flat fairness–performance frontier in this task. The fairness penalty does not rescue a weak model by sacrificing prediction accuracy. Rather, it reduces disparities while retaining the predictive gains of nonlinear structure. This empirical shape is important, because it suggests that material fairness improvements are available without a meaningful collapse in underwriting quality.

6 Interpretability Results

Aggregate metrics alone do not establish that the resulting system is suitable for regulated lending. The model also has to remain intelligible. FAN's additive form makes it possible to examine what the model learns, how fairness regularization changes its logic, and whether the resulting explanation structure is usable for adverse-action communication.

6.1 Predictor importance and shape functions

Figure 1 reports the most important model terms under the unconstrained and fairness-aware specifications, as measured by the mean absolute output of each shape function over the training data. Across both specifications, the top predictors are dominated by univariate effects. For $\lambda_{\text{EO}} = 0$, the two largest contributors are `loan_purpose_1` and `dti_num`, followed by `log1p(loan_amount)` and `loan_to_value_ratio`. The leading interactions enter next: `log1p(loan_amount)×log1p(property_value)` and `dti_num×loan_type_1`, along with contextual terms such as `tract_minority_population_percent`.

With fairness regularization ($\lambda_{\text{EO}} = 3$), 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. Notably, some geographic proxies lose relevance under the penalty: `tract_minority_population_percent` drops out of the top set, while `log1p(income)` enters the top fifteen, reflecting a modest shift toward direct repayment capacity measures. The fairness gain therefore comes from changing the composition of the decision rule, not from deleting substantive underwriting content.

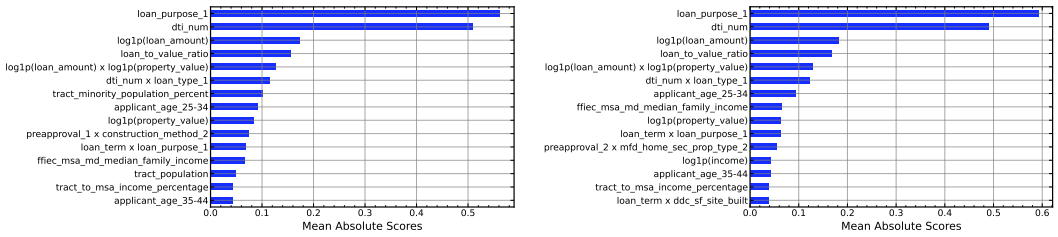


Figure 1: Global term importance. Fairness regularization preserves the prominence of core underwriting variables while reducing the influence of some proxy-like geographic terms.

Figure 2 presents the shape functions f_i for the four most important univariate predictors, comparing the unconstrained model ($\lambda_{\text{EO}} = 0$) with the penalized model ($\lambda_{\text{EO}} = 3$). The effect of `loan_purpose_1` decreases the model output for home-purchase loans while non-purchase purposes are penalized—consistent with refinances and cash-out loans raising borrower leverage and collateral risk. The debt-to-income ratio produces a monotonically increasing effect on denial scores, with a sharper increase at high values reflecting nonlinear risk once borrowers cross common underwriting thresholds. Loan amount is distinctly non-monotone, penalizing both very small and very large loans relative to the middle of the distribution, capturing a U-shaped risk profile. The loan-to-value ratio is approximately flat through ordinary ranges and then rises sharply once leverage becomes high, consistent with low borrower equity raising expected losses.

The comparison between the two models illustrates how the predictive process adjusts for the fairness objective. Fairness regularization raises rejection probabilities for non-home-purchase loans (suggesting purpose correlates with protected-group status), dampens the impact of high DTI on the classification output, and reallocates effects at loan-amount extremes. For LTV, the EO penalty leaves the functional form essentially unchanged—the fairness objective does not rewrite underwrit-

ing logic, but tilts certain margins in ways consistent with reducing EO disparities while leaving core risk gradients intact.

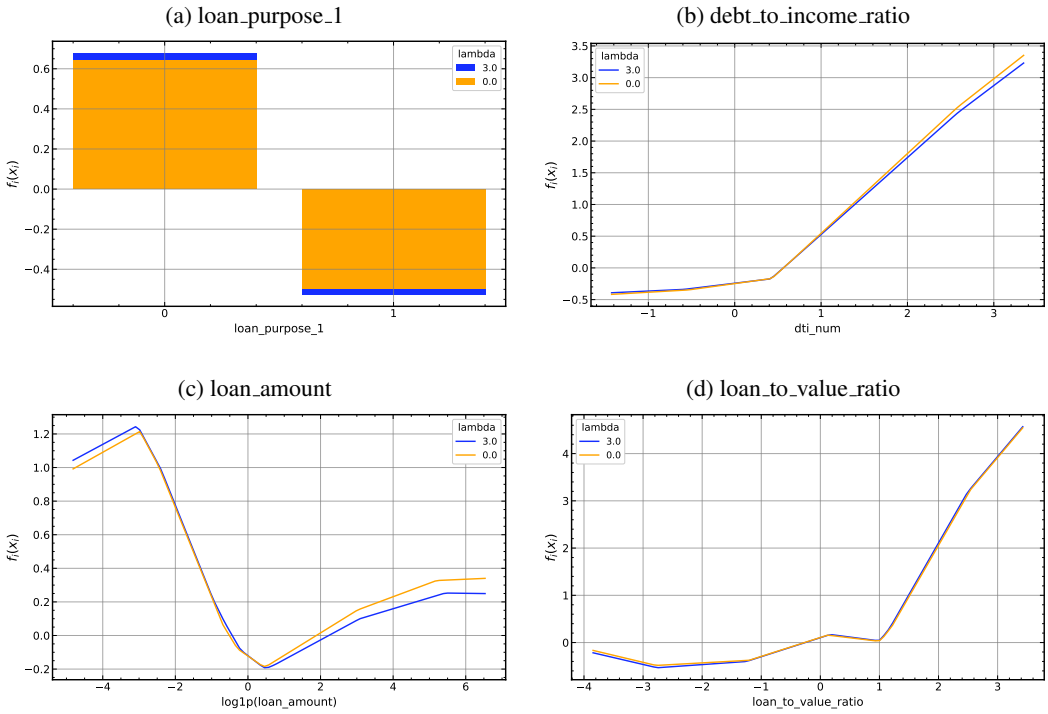


Figure 2: Univariate shape functions $f_i(x_i)$ for the four most important predictors, comparing $\lambda_{\text{EO}} = 0$ (blue) and $\lambda_{\text{EO}} = 3$ (orange).

6.2 Adverse-action explanations

For lending applications, interpretability is not only about global understanding. A model must also be able to explain individual adverse decisions. Because FAN scores are additive, the same decomposition used for global inspection can be used to summarize which terms push rejected applications upward toward denial. For any main effect or interaction term t, let $f_t(X_l)$ denote its additive contribution to the score $s_\theta(X_l)$. To summarize adverse-action drivers, we focus on rejected applications, $\mathcal{D}_1 = \{l : Y_l = 1\}$, and define the mean positive score $S_t^+ = \frac{1}{|\mathcal{D}_1|} \sum_{l \in \mathcal{D}_1} \max\{f_t(X_l), 0\}$. S_t^+ measures how strongly term t pushes denied applications upward toward rejection on average.

We find that, 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` (see also Figure 6 in the Appendix for a summary of the biggest contributors). 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. It remains 2.22 under $\lambda_{\text{EO}} = 3$. The interaction `dti_num×loan_type_1` is also material at about 0.26 in both cases, suggesting that repayment-capacity constraints matter differently across conventional versus insured or guaranteed channels. Programmatic and size effects follow: `loan_purpose_1` contributes 0.64 in the unconstrained model and 0.68 under fairness regularization, while logged loan amount remains around 0.40.

Under fairness regularization, the model preserves the same broad underwriting logic while modestly reallocating adverse-action mass across terms. The preapproval-related interaction that appears at $\lambda = 0$ drops out of the top contributors at $\lambda = 3$, suggesting reduced reliance on process-related interactions. The main denial drivers, however, remain consistent: DTI, loan purpose, loan amount, and the DTI–loan-type interaction continue to account for most upward movements in the denial score. At the same time, the entry of `applicant_age_8888` under $\lambda = 3$ indicates that missing

or withheld age information may capture informational frictions or atypical borrower characteristics. Thus, fairness regularization changes the composition of adverse-action contributions without replacing the model's core underwriting logic.

6.3 Decomposition of the EO disparity

The model's additive structure also allows us to decompose the equalized-odds disparity near the operating margin. For each outcome class $y \in \{0, 1\}$ and group $a \in \{0, 1\}$, define $\bar{f}_t^{a,y} = \mathbb{E}[f_t(X) \mid A = a, Y = y]$ and $\Delta_t^{(y)} = \bar{f}_t^{1,y} - \bar{f}_t^{0,y}$. Because the FAN score is additive, these term-level differences exactly sum to the class-conditional minority–non-minority score gap:

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

The quantities $\Delta_t^{(y)}$ therefore identify which features or interactions generate the EO-related score gap, and how this composition changes with λ_{EO}.

The results indicate that collateral and program/product variables consistently contribute at high to moderate magnitudes with stable signs (see also Figure 5 in the Appendix which presents the ten largest $\Delta_t^{(y)}$ across outcomes and penalization levels). `loan_purpose_1` (home purchase) is the single largest negative contributor near the threshold for both classes and both λ values (from -0.200 to -0.153 in denials), implying that, conditional on outcome, purchase purpose lowers the average score for minority applicants relative to non-minorities. Repayment capacity remains a primary driver: `dti_num` is large and positive at $\lambda = 0$ (0.111 in denials, 0.167 in originations) and changes only modestly under fairness regularization (0.105 and 0.174, respectively).

In contrast, geographic proxies compress substantially. `tract_minority_population_percent` contributes 0.131 to the denial-side score gap and 0.147 to the origination-side gap at $\lambda = 0$, and falls to 0.038 and 0.041 under $\lambda_{\text{EO}} = 3$. The fairness penalty is therefore not mainly shrinking disparities by weakening the model's attention to repayment capacity. It is shrinking them by reducing the role of variables that are more plausibly functioning as proxies for protected status or neighborhood disadvantage, while preserving central underwriting signals.

6.4 Interactions

Moreover, the interaction surfaces reveal how FAN preserves economically recognizable underwriting logic. Figure 3 shows two representative pairwise effects. The first interaction links loan amount to property value. In both the unconstrained and fairness-aware models, the surface is close to zero for small loans secured by relatively valuable collateral, but becomes sharply positive as the requested loan grows large relative to the property's value. The learned interaction therefore behaves like an additional leverage screen layered on top of the separate loan-amount and collateral effects. Importantly, fairness regularization does not remove this structure; if anything, it makes the ridge more concentrated in the riskiest corner, which is consistent with preserving a strong penalty for thinly collateralized borrowing.

The second interaction combines DTI with conventional-loan status. For conventional loans, moderately high DTI raises the rejection score sharply. For insured or guaranteed loans, the same DTI range is treated more leniently, consistent with the fact that FHA, VA, or USDA channels partially insure lenders against default risk. Under fairness regularization the overall pattern remains intact, but the model becomes somewhat more favorable to the insured/guaranteed segment in the range where minority applicants are plausibly overrepresented. This is useful substantively because it shows that fairness gains do not come from flattening all nonlinear structure. They come from rebalancing how the model uses product channels and repayment-capacity signals.

7 Analysis of model behavior

So far, our empirical results have shown sizable minority-specific disparities in error rates and suggested that fairness regularization can shrink those gaps while retaining overall model performance. We now ask: do these results reflect real differences in how similar applicants are treated at the actual decision margin and how can these effects be understood?

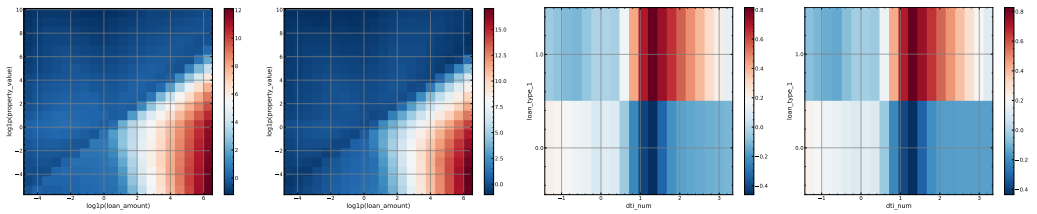


Figure 3: Representative interaction effects. Fairness regularization preserves the economically central pairwise structure of the model rather than collapsing the score into a purely additive or nearly linear rule.

Cutoff behavior. We leverage a discontinuity at the LTV$= 80\%$ underwriting threshold—a well-established boundary for requiring private mortgage insurance (PMI)—to implement a sharp regression discontinuity (RD) design. For loan application l, let the running variable be $r_l := \text{LTV}_l - 80$, and let $D_l := \mathbf{1}\{r_l \geq 0\}$ indicate being weakly above the threshold. On the test data, we estimate a local linear RD with triangular kernel weights, comparing model-implied approvals across the boundary (Green & Wachter, 2005; Adelino et al., 2025).

Table 2 reports estimates for model-generated approvals with bandwidth $h = 10$ percentage points. We find a sizable and precisely estimated negative jump in approval probability at LTV$\approx$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, $\hat{\zeta}$, is large and negative across all specifications. Crucially, $\hat{\zeta}$ becomes less negative as the fairness weight increases, from -7.74 p.p. at $\lambda = 0$ to -6.86 p.p. at $\lambda = 3$. The same pattern holds at bandwidths $h \in \{3, 5\}$. The model preserves these institutional cutoffs while making their minority shortfalls smaller. Figure 4 illustrates the discontinuity.

Table 2: RD estimates of $Y_l = \alpha + \beta\, r_l + \gamma\, D_l + \theta\, A_l + \zeta\,(A_l \times D_l) + \eta\,(D_l \times r_l) + \varepsilon_l$ around LTV$= 80\%$ with bandwidth $h = 10$ p.p. $\hat{\gamma}$ is the non-minority jump; $\hat{\zeta}$ is the minority differential. Standard errors are in parentheses.

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

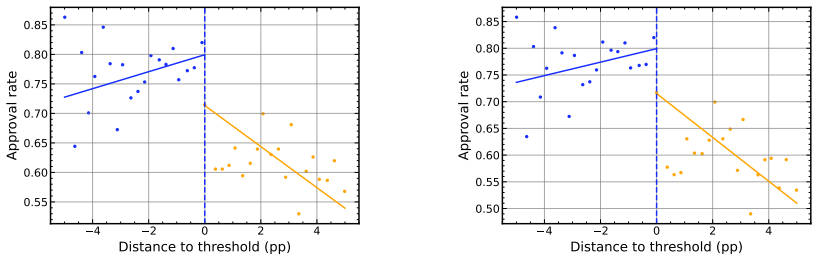


Figure 4: Validation at an institutional underwriting cutoff. The fairness-aware model preserves the LTV$= 80\%$ discontinuity while reducing the minority-specific shortfall at that boundary.

Within-lender fairness. We next ask whether the fairness gains come from changing applicant ordering *within* lenders or only from shifting composition *across* lenders. To address this question, we hold each lender's approval capacity fixed and recompute counterfactual model decisions. Table 3 shows that the volume-weighted mean within-lender TPR gap falls from 0.0804 to 0.0408, and the within-lender FPR gap falls from 0.0419 to 0.0142. An Oaxaca–Blinder-style decomposition (Oaxaca, 1973; Blinder, 1973) shows that between-lender components are negligible, so the gains come primarily from changes in ranking and decision logic within institutions rather than from lender reallocation. This suggests that lenders could meaningfully reduce disparities by modifying their internal scoring rules, even without changing pricing, capacity, or market structure.

Table 3: Within-lender, capacity-matched fairness by model. Weighted means summarize lender-level gaps; decomposition columns separate within-lender from between-lender contributions.

Model	Mean ΔTPR_ℓ	Mean ΔFPR_ℓ	Within ΔTPR	Between ΔTPR	Within ΔFPR	Between ΔFPR
FAN ($\lambda = 0$)	0.0804	0.0419	0.0757	-0.0001	0.0433	-0.0002
FAN ($\lambda = 1$)	0.0647	0.0298	0.0616	-0.0001	0.0301	-0.0001
FAN ($\lambda = 2$)	0.0519	0.0229	0.0488	-0.0012	0.0230	-0.0001
FAN ($\lambda = 3$)	0.0408	0.0142	0.0400	-0.0005	0.0147	-0.0003

Approved-loan risk and pricing alignment. A final concern is whether fairness regularization creates distortions in approved-loan pricing. We run separate regressions of the form $\text{price}_l = \alpha + \beta\, s_\theta(X_l) + \gamma\, \mathbf{1}\{A_l{=}1\} + \varepsilon_l$ for the HMDA pricing variables interest rate and rate spread. Among approved loans, the minority gap in modeled rejection risk falls from 2.32 to 1.14 percentage points, but pricing relationships remain effectively unchanged. The conditional minority coefficient γ in an interest-rate regression stays around 0.0065–0.0068 percentage points, and the corresponding rate-spread coefficient remains around 0.011–0.012 (Table 7 of Appendix E summarizes the full results). Overall, we find no evidence that fairness regularization induces pricing asymmetries or cross-subsidization within approved loans.

Additionally, we conduct several robustness tests confirming our main results (see the Appendix for all details). Taken together, these exercises show that the fairness-aware model behaves well where institutions make decisions: at underwriting cutoffs, inside lenders, and among approved loans for which pricing information is observed. The DTI$= 43\%$ boundary shows the same qualitative pattern as LTV$= 80\%$, and placebo tests at LTV$= 70\%$ produce small, statistically insignificant results, tying the main cutoff findings to institutional boundaries. Additionally, we find that the main results replicate in 2019 HMDA data, with even stronger fairness improvements: the TPR gap falls from 0.047 to 0.028 and the FPR gap from 0.009 to 0.003 as λ_{EO} rises from 0 to 3, addressing the possible concern that the 2018 findings are idiosyncratic to one year.

8 Conclusion

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.

Our findings have two main 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 for fairness-aware alternatives. Second, from a research perspective, the relatively flat fit-fairness tradeoff 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.

AI USE STATEMENT

In this work, we used generative AI tools to assist with editing portions of the manuscript, including improving clarity and concision. All AI-assisted text was reviewed and revised by the authors, who verified the accuracy of the technical content, claims, and citations. We take responsibility for the final content of this work, including text produced or edited with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide additional information for reproducibility of the main results in the paper and appendix. Appendix D describes the FAN architecture, training procedure, regularization, and interaction selection, while Section 4 and Appendix B document sample construction, variable definitions, and preprocessing of the HMDA data. Formal statements and proofs for the theoretical properties of the fairness objective are provided in Appendix G. The main text also reports the data splits, evaluation metrics, fairness-penalty values, and experimental comparisons used in the reported results.

REFERENCES

Manuel Adelino, Antoinette Schoar, and Felipe Severino. Credit supply and house prices: Evidence from mortgage market segmentation. *Journal of Financial Economics*, 163:103958, January 2025. ISSN 0304-405X. doi: 10.1016/j.jfineco.2024.103958.

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

Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics*, 143(1):30–56, 2022.

Patrick Bayer, Fernando Ferreira, and Stephen L. Ross. What drives racial and ethnic differences in high-cost mortgages? The role of high-risk lenders. *Review of Financial Studies*, 31(1):175–205, 2018.

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

Tobias Berg, Valentin Burg, Ana Gombović, and Manju Puri. On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies*, 33(7):2845–2897, 2020.

Emily Black, John Logan Koepke, Pauline T. Kim, Solon Barocas, and Mingwei Hsu. Less discriminatory algorithms. *Georgetown Law Journal*, 113(1):53–120, 2024.

Alan S. Blinder. Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4):436–455, 1973.

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

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

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

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, New York, 2 edition, 1999. ISBN 978-0-471-31716-6.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1):5–47, 2022.

Andreas Fuster, Stephanie H. Lo, and Paul S. Willen. The time-varying price of financial intermediation in the mortgage market. *The Journal of Finance*, 79(4):2553–2602, 2024. ISSN 1540-6261. doi: 10.1111/jofi.13358.

Kristopher Gerardi, Paul S. Willen, and David Hao Zhang. Mortgage prepayment, race, and monetary policy. *Journal of Financial Economics*, 147(3):498–524, March 2023. ISSN 0304-405X. doi: 10.1016/j.jfineco.2022.12.001.

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

Richard K. Green and Susan M. Wachter. The American mortgage in historical and international context. *Journal of Economic Perspectives*, 19(4):93–114, 2005.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1, Part B):429–450, May 2021. ISSN 0304-4076. doi: 10.1016/j.jeconom.2020.07.009.

Moritz Hardt, Eric Price, and Nati Srebro. 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, pp. 3315–3323, 2016.

Sabrina T. Howell, Theresa Kuchler, David Snitkoff, Johannes Stroebel, and Jun Wong. Lender automation and racial disparities in credit access. *Journal of Finance*, 79(2):1457–1512, 2024.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 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)*, pp. 431–432. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

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

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 623–631, New York, NY, USA, August 2013. Association for Computing Machinery. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487579.

Colin McDiarmid. 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, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, 1973.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x.

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

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019. ISSN 1533-7928.

APPENDIX

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 4: 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 5: 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.

Continued on next page

Abbreviation	HMDA Variable	Values / HMDA Codes
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 – Multi-family: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
ffiec_msa_md_median_family_income	FFIEC Median family income of the MSA/MD[1]	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
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

Continued on next page

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

Abbreviation	HMDA Variable	Values / HMDA Codes
`tract_to_msa_income_percentage`	Tract-to-MSA Income Percentage	Varying values

Table 6: 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. 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

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

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

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

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

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

yield the decomposition

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

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 $q = \hat{p}_\theta = p^\star$.[2] 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)$=R$.[3]

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 during training by replacing d_θ with

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

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

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

[2] 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 $\tfrac{p^\star}{q^2}+\tfrac{1-p^\star}{(1-q)^2}>0$ ensures a unique minimum.

[3] 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.

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., $G = J_{\text{fair}}^{1/2}$) and $\varepsilon \ge 0$ is a tolerance. The corresponding Lagrangian is

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

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

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

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 $J_{\text{fair}}(\theta)$ penalizes disparities expressed in terms of approval or error rates that also determine profit via $R\cdot\text{TP} - C\cdot\text{FP}$. 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

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

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 $i \in \mathcal{C}$ 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 3.2 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),

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

  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:

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

  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

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

  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 $\mathrm{Var}[v_{ij} f_{ij}(\tilde{x}_i, \tilde{x}_j)]$ 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 $\widehat{\mathcal{P}}_{\text{clar}}(\theta)$ 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

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

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:

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

where $g_i : \{0, \ldots, B_i - 1\} \to \mathbb{R}$ and $g_j : \{0, \ldots, B_j - 1\} \to \mathbb{R}$ 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

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

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

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

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

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 $\text{scale}_{ij} = |v_{ij}|^2 \cdot \text{Var}[f_{ij}(\tilde{x}_i, \tilde{x}_j)]$ 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 $q\%$. 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 ADDITIONAL TABLES AND FIGURES

Table 7: Approved-loan parity and pricing alignment. The last two columns report the conditional minority coefficient from pricing regressions among approved loans.

Model	$\Delta\bar{p}$	Minority coef. (rate)	Minority coef. (spread)
FAN ($\lambda = 0$)	0.0232	0.0068	0.0111
FAN ($\lambda = 1$)	0.0172	0.0067	0.0113
FAN ($\lambda = 2$)	0.0131	0.0065	0.0119
FAN ($\lambda = 3$)	0.0114	0.0066	0.0114

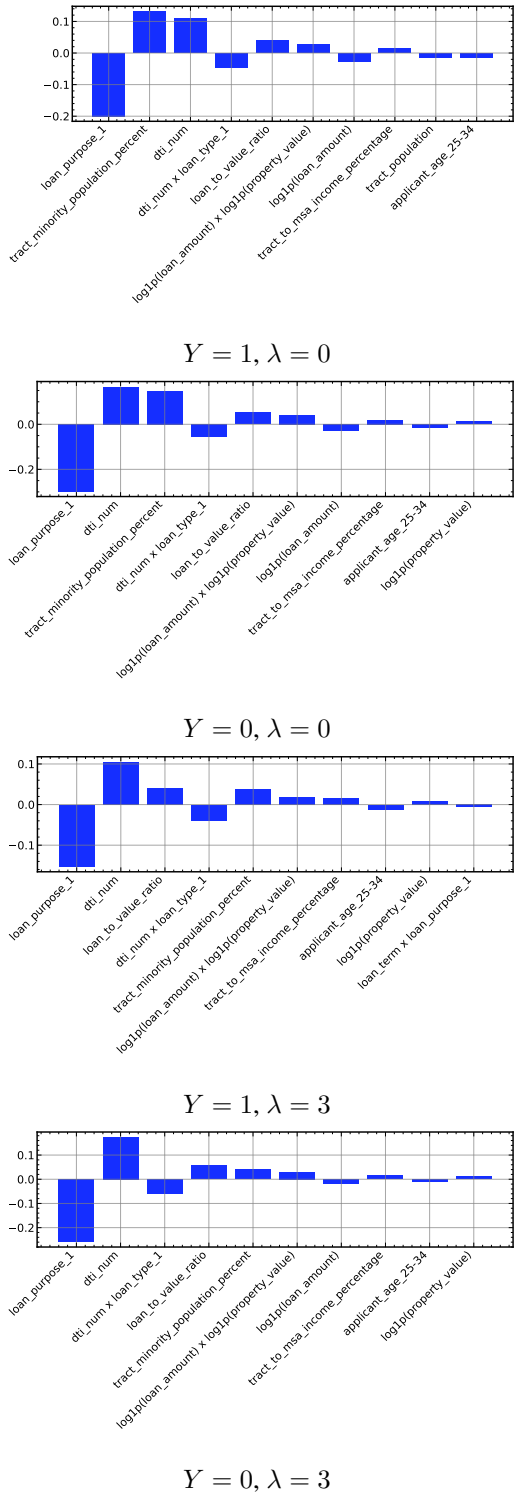


Figure 5: $\Delta_t^{(y)}$ near the decision boundary. The ten largest per-term contributions to the cross-group score gap, for each class y and fairness weight λ. Fairness regularization reduces dependence on geographic correlates while preserving economically grounded channels.

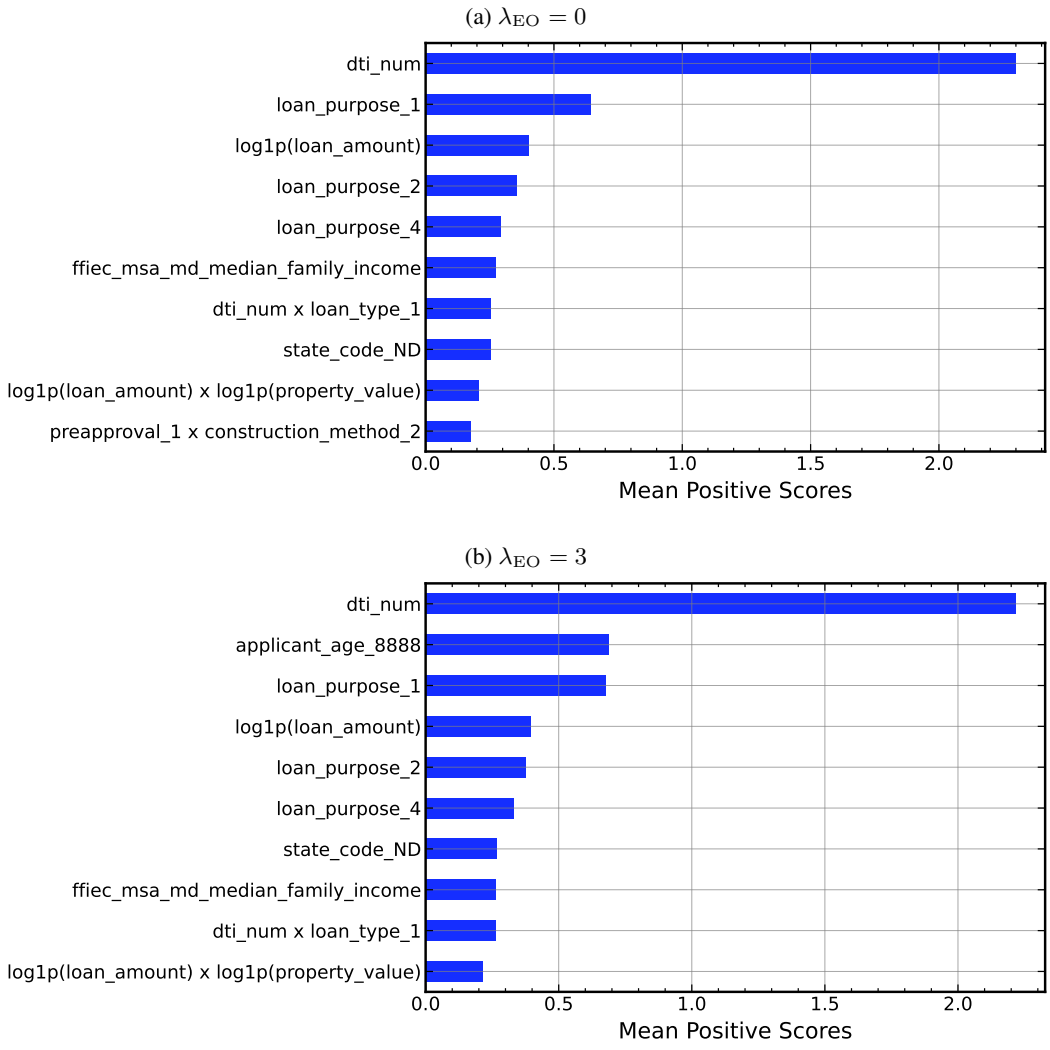


Figure 6: Mean positive scores S^+ (top ten). This figure compares the most important contributors to rejections for the FAN models with $\lambda_{\text{EO}} = 0$ (Panel A) and $\lambda_{\text{EO}} = 3$ (Panel B).

F PROPERTIES OF THE FAIRNESS PENALTY

Theoretical analysis shows multiple desirable properties of the penalty term (see the full proofs in the subsequent Appendix). Specifically, we first show that the smooth penalty is most consequential near the decision boundary, which is precisely where lending decisions flip and where fairness interventions matter most for applicants. Let

$$\mu_{g,y}(\theta) = \mathbb{E}[\sigma(s_\theta(X)) \mid A = g, Y = y]$$

denote the average predicted denial probability within group g and outcome class y. If we define the threshold-indexed rate curve

$$R_{g,y}(\tau) = \Pr\big(s_\theta(X) \geq \tau \mid A = g, Y = y\big),$$

then we can show that

$$\mu_{g,y}(\theta) = \int_{-\infty}^{\infty} \sigma'(\tau) R_{g,y}(\tau)\, d\tau,$$

so the fairness penalty can be read as a weighted average of group differences in TPR and FPR curves over thresholds, with weight $\sigma'(\tau)$ concentrated near the decision boundary. In other words, training is threshold-free, but the fairness signal is not arbitrary: it is largest in the score region where small changes in the model actually flip approval decisions. This provides a direct connection between the soft, threshold-agnostic penalty formulation and fairness outcomes that matter when applying the model in practice, which requires a hard threshold τ (we use $\tau = 0$) that turns the continuous model output into a binary decision.

The same analysis also provides an interpretable bound. Let $\mathcal{P}_{\text{mean}}$ denote the sum of the group variances of $\mu_{g,1}$ and $\mu_{g,0}$, and let $\mathcal{D}_1 + \mathcal{D}_0$ denote the threshold-averaged equalized-odds dispersion in the rejection and origination classes. Then the theoretical results imply

$$\mathcal{P}_{\text{mean}}(\theta) \leq \mathcal{D}_1(\theta) + \mathcal{D}_0(\theta).$$

This does not mean that every pointwise EO gap must be small whenever the penalty is small, but it does show that reducing the training penalty controls a decision-relevant average of cross-group error-rate disparities. Moreover, we establish an out-of-sample generalization guarantee of the usual $O(1/\sqrt{n})$ form, scaled by the smallest group-outcome cell mass and by model-complexity quantities such as sparsity and smoothness. The key implication is that the fairness term is not merely a computational trick or an in-sample artifact: it is an auditable surrogate with a transparent relationship to the operational notion of fairness that matters in lending.

G 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 these properties mathematically by formulating three propositions:

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.

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,*

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

Consequently,

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

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

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

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.*

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.[4]

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 $w(\tau) \geq 0$ 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.

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

Application in Proposition 1. We apply Theorem 1 with

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

so that

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

G.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

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

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

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

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

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

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

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

G.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 $w(\tau) \geq 0$ and $\int w = 1$, $w(\tau)\, d\tau$ is a probability measure. By Jensen's inequality for the convex map $z \mapsto z^2$,

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

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

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

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

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

Adding these inequalities and using linearity of the integral yields

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

Therefore:

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

□

G.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

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

with base functions satisfying $\|f_i\|_\infty \leq B_1$ (values never exceed B_1), $\mathrm{Lip}(f_i) \leq L_1$ (slope $\leq L_1$), $\|f_{ij}\|_\infty \leq B_2$, $\mathrm{Lip}(f_{ij}) \leq L_2$. 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:

$$\Pr(A{=}g,\, Y{=}y) \geq p_{\min} > 0 \quad \text{for all } (g, y) \in \mathcal{G} \times \{0, 1\}. \tag{4}$$

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

$$\sup_{z_1, \ldots, z_m, z_i'} \big| F(z_1, \ldots, z_i, \ldots, z_m) - F(z_1, \ldots, z_i', \ldots, z_m) \big| \;\leq\; c_i \quad \textit{for each } i.$$

Then, for all $t > 0$,

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

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.[5] 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

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

and let $(X'_1, \ldots, X'_{n_{g,y}})$ 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 "$\le$".)

Now introduce Rademacher signs $\{\varepsilon_\ell\}_{\ell\in I_{g,y}}$, i.i.d. with $\varepsilon_\ell \in \{\pm 1\}$ 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 $I_{g,y}$),

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

Since X'_ℓ and X_ℓ have the same distribution,

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

[5] 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).

Combining the displays yields exactly the desired inequality:

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

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

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

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 $I_{g,y}$),

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

(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 $1-\delta/2$

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

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

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

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.[6] Then, for each realization of ε,

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

[6]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.

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

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

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

$$\widehat{\mathfrak{R}}_m(\mathcal{S}) \;=\; \Lambda\,\mathbb{E}_\varepsilon\left[\left\|\frac{1}{m}\sum_{\ell=1}^m \varepsilon_\ell\,\phi(x_\ell)\right\|_\infty\right] \;=\; \Lambda\,\mathbb{E}_\varepsilon\left[\max_{1\le j\le M}\left|\frac{1}{m}\sum_{\ell=1}^m \varepsilon_\ell\,\phi_j(x_\ell)\right|\right]. \tag{5}$$

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 $t > 0$,

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

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

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

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

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

Substituting this bound into equation 5 gives, for the block,

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

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

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

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 $a_g, b_g \in [0, 1]$, $1 \leq g \leq G$. Then*

$$\Big|\,\mathrm{Var}_g(a_g) - \mathrm{Var}_g(b_g)\Big| \;\leq\; \frac{4}{G}\sum_{g=1}^{G} |a_g - b_g|.$$

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

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

Additionally,

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

Adding the two previous results gives:

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

□

Theorem 3 (Alternative formulation of Proposition 3)**.** *Under equation 4, 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 $y \in \{0, 1\}$,

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

Summing over the two labels y gives

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

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

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

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 $n_{g,y} = \sum_{\ell=1}^{n} Z_\ell^{(g,y)} \sim \text{Binomial}(n, p_{g,y})$ with $p_{g,y} = \Pr(A{=}g, Y{=}y) \ge p_{\min}$ by equation 4. The multiplicative Chernoff bound gives, for any $\eta \in (0, 1)$,

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

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

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

On the event

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

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

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

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 $1 - \delta$.

□

H 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 3.2, 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 8, 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 8 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 8: Penalty term and average (threshold-weighted) group dispersion. This table evaluates penalty and dispersion terms across data sets and models.

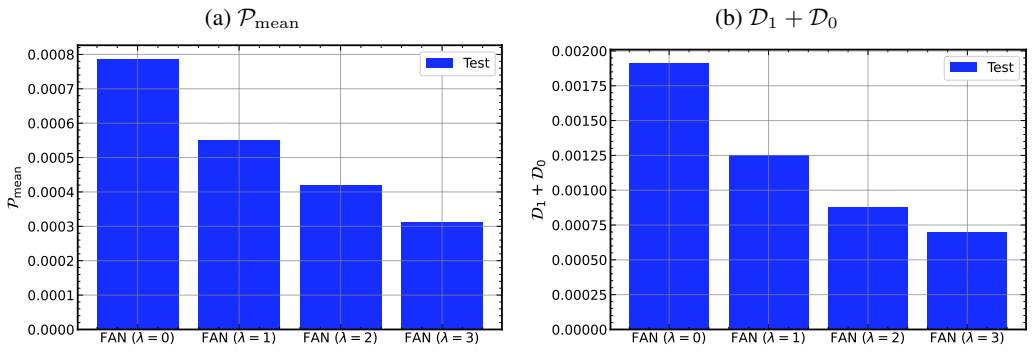


Figure 7: Empirical penalty and upper bound. This figure summarizes the out-of-sample mean penalty (Panel A) and total dispersion bound (Panel B) across penalization levels.

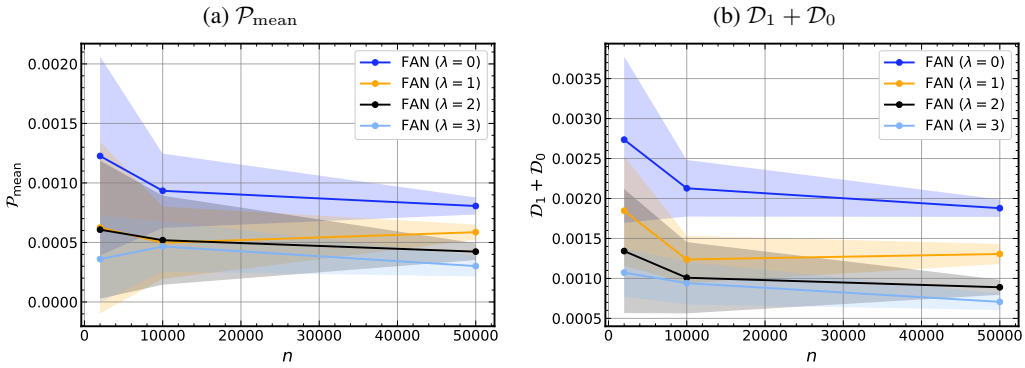


Figure 8: Learning curves. This table reports the mean fairness penalty (Panel A) and dispersion boundary (Panel B) computed over hold-out test sets of varying sizes n. Shaded regions indicate the one standard deviation bands around the mean.

I 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 9 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 10 and 11, 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	$\Delta\text{TPR}(\tau^\star)$	$\lvert\Delta\text{TPR}(\tau^\star)\rvert$	$\Delta\text{FPR}(\tau^\star)$	$\lvert\Delta\text{FPR}(\tau^\star)\rvert$	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 9: Fairness summary for observed rate-matched decision thresholds. This table reports performance and fairness metrics across models with matched decision thresholds.

τ	Appr. rate	TPR_0	TPR_1	$\Delta\mathrm{TPR}$	$\lvert\Delta\mathrm{TPR}\rvert$	FPR_0	FPR_1	$\Delta\mathrm{FPR}$	$\lvert\Delta\mathrm{FPR}\rvert$
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 10: Metrics at varying thresholds ($\lambda_{EO} = 0$)

τ	Appr. rate	TPR_0	TPR_1	$\Delta\mathrm{TPR}$	$\lvert\Delta\mathrm{TPR}\rvert$	FPR_0	FPR_1	$\Delta\mathrm{FPR}$	$\lvert\Delta\mathrm{FPR}\rvert$
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 11: Metrics at varying thresholds ($\lambda_{EO} = 3$)

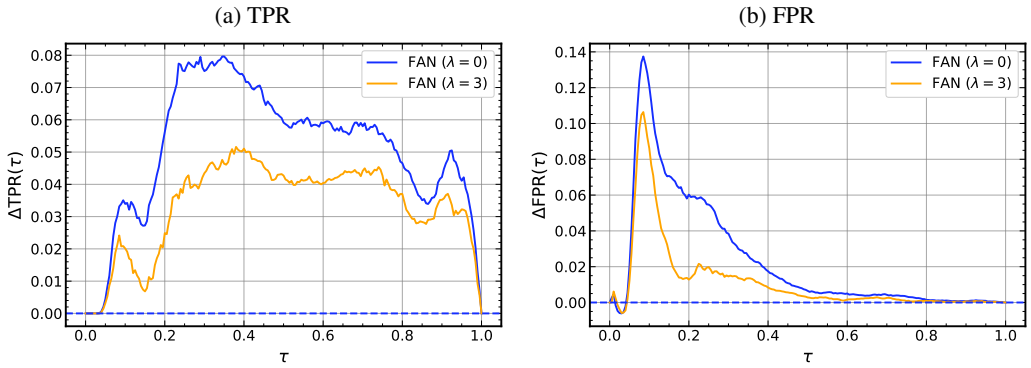


Figure 9: Rate disparities across thresholds. This figure presents group differences in true positive rates (Panel A) and false positive rates (Panel B) across decision thresholds, comparing the unconstrained and the penalized classifier.

J ROBUSTNESS: IDENTIFICATION

J.1 IDENTIFICATION 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.[7] 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. (6) 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:

$$Y_l \;=\; \alpha + \beta\, r_l + \gamma\, D_l + \theta\, A_l + \zeta\,(A_l \times D_l) \;+\; \eta\,(D_l \times r_l) \;+\; \varepsilon_l, \tag{6}$$

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 12 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 10 presents local linear fits on each side of the cutoff, illustrating the discontinuity.[8] Table 13 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 7 that our fairness regularization successfully reduces minority-specific approval gaps at the decision boundary.

[7] 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.

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

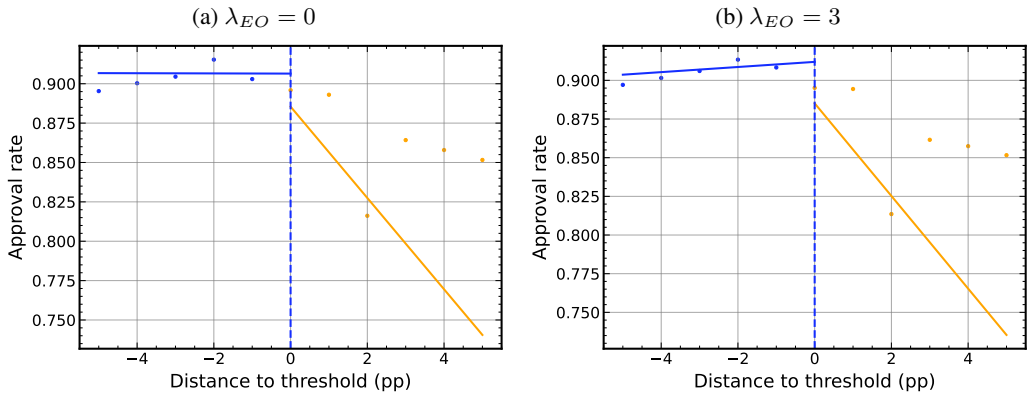


Figure 10: Binned means and local linear fit at DTI=43% for model approvals (h=5). The figure shows loan-to-value ratios and approval rates for the unconstrained FAN (Panel A) and the penalized model (Panel B) with the corresponding regression lines below and above the 80% threshold.

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

Table 12: Regression results around DTI= 43% (bandwidth h=10 p.p.). This table reports intercepts and discontinuities estimated from the RD design specified in (6). 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 13: 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 (6) 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: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

J.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 14 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 RD identification strategy 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 14: 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 (6) 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: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

J.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 11 and Table 16) and the LTV cutoff (Figure 12 and Table 15).

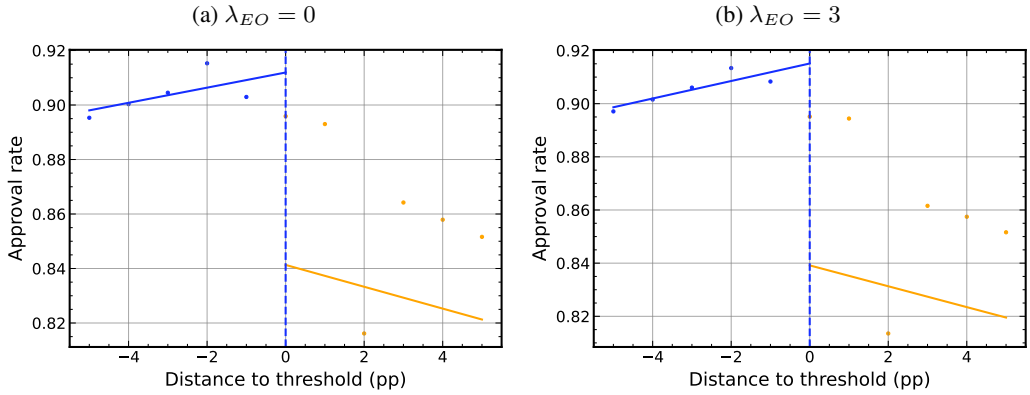


Figure 11: Local linear fit at DTI=43% for model approvals (uniform kernel)

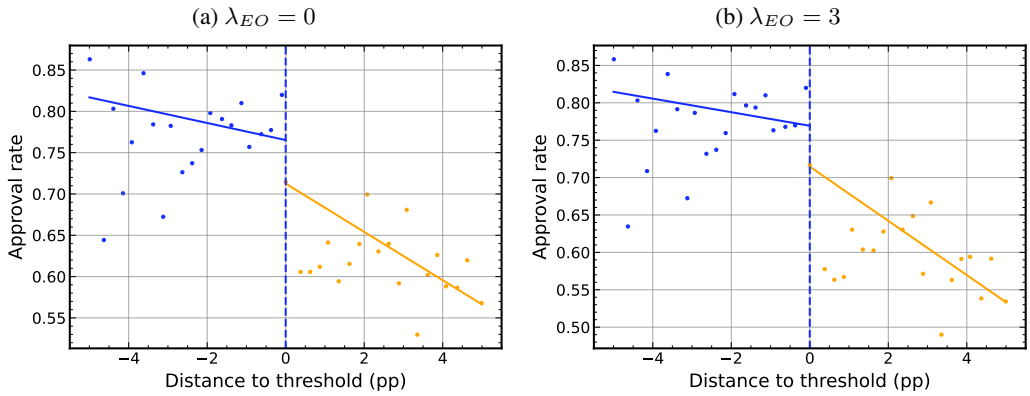


Figure 12: Local linear fit at LTV=80% for model approvals (uniform kernel)

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 15: 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 16: Local linear fit at DTI=43% across bandwidths $h \in \{3, 5, 10\}$ (uniform kernel)

K 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 17 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 $\lambda_{\mathrm{EO}} = 0$ to 0.043 at $\lambda_{\mathrm{EO}} = 3$. 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	$\lvert\mathrm{TPR}_1 - \mathrm{TPR}_0\rvert$	$\lvert\mathrm{FPR}_1 - \mathrm{FPR}_0\rvert$
Logit (baseline)	0.829	0.801	0.064	0.012
FAN ($\lambda_{\mathrm{EO}} = 0$)	0.838	0.821	0.058	0.007
FAN ($\lambda_{\mathrm{EO}} = 1$)	0.839	0.822	0.051	0.004
FAN ($\lambda_{\mathrm{EO}} = 2$)	0.839	0.822	0.045	0.004
FAN ($\lambda_{\mathrm{EO}} = 3$)	0.839	0.822	0.043	0.003

Table 17: 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 ($a = 0$).

L ROBUSTNESS: ALTERNATIVE DATA

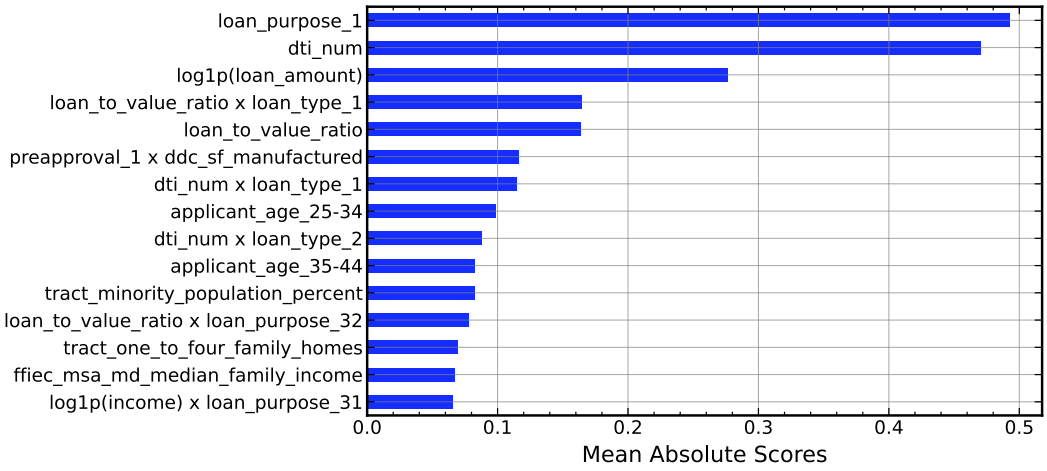
As a test of robustness, we repeat our analysis on 2019 HMDA data. Table 18 shows that the 2019 results closely mirror (and in several respects strengthen) the main 2018 findings. Predictive performance is stable or slightly higher in 2019: the baseline Logit accuracy rises from 0.829 (2018) to 0.861 (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 13). 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 14). 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 18: 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 ($a = 0$).

(a) $\lambda_{EO} = 0$



(b) $\lambda_{EO} = 3$

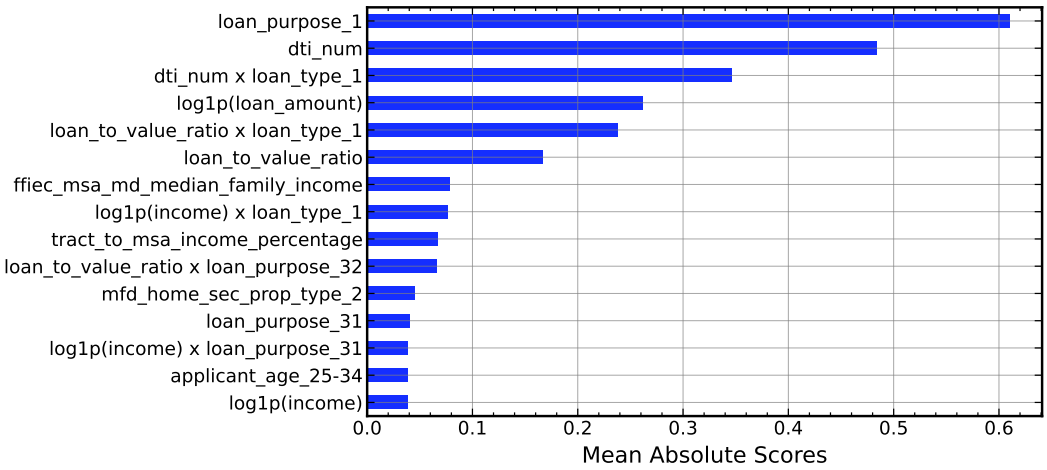


Figure 13: Mean Absolute Scores $S(i)$ and $S(i, j)$

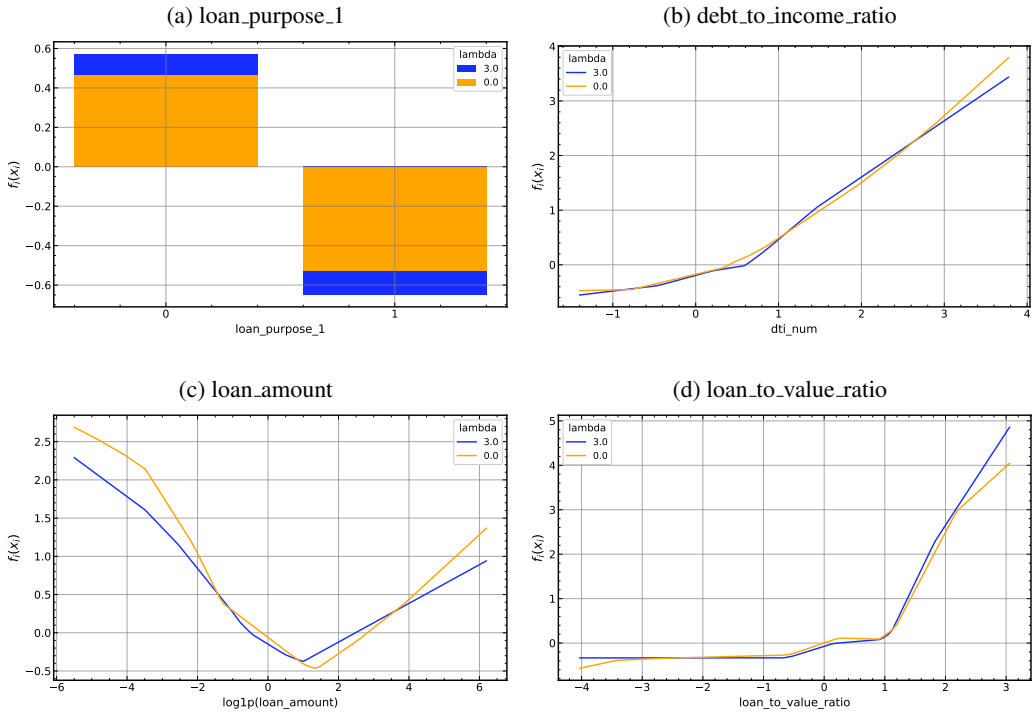


Figure 14: Univariate shape functions $f_i(x_i)$

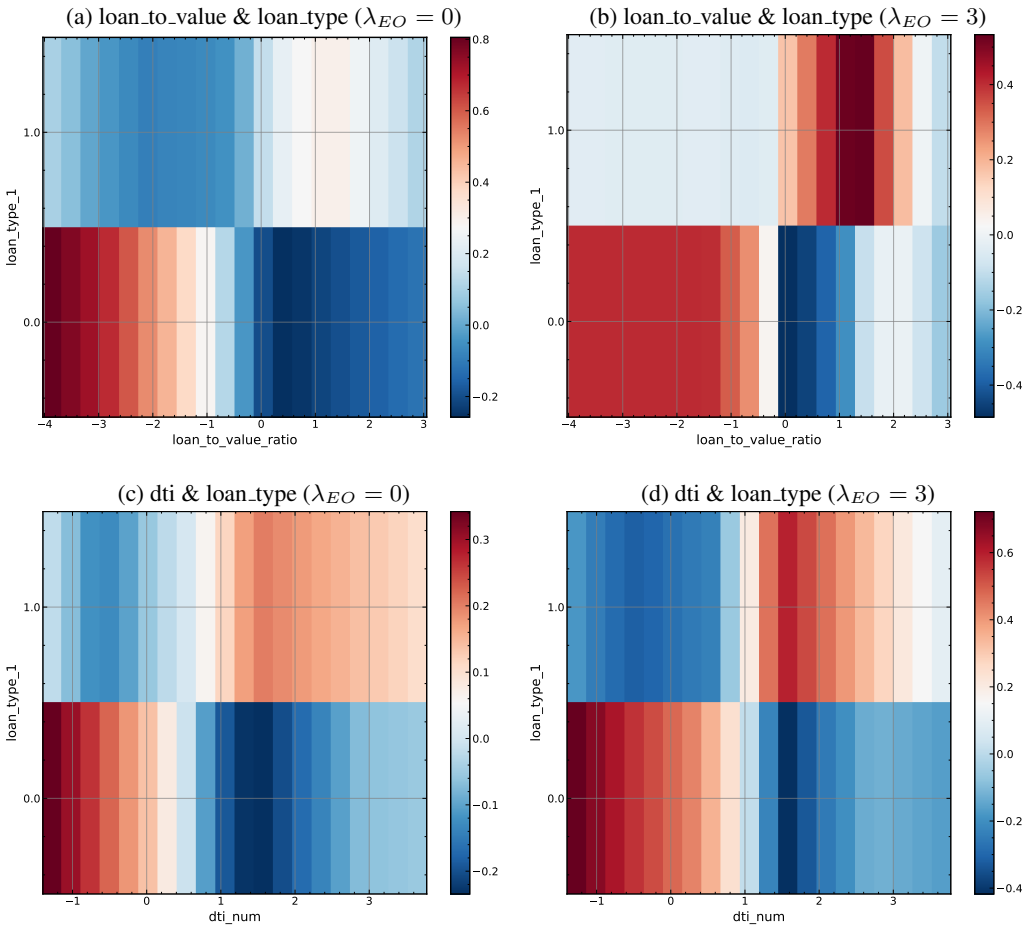


Figure 15: Interaction effects $f_{ij}(x_i, x_j)$

M 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 19 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	$\lvert\mathrm{TPR}_1 - \mathrm{TPR}_0\rvert$	$\lvert\mathrm{FPR}_1 - \mathrm{FPR}_0\rvert$
Logit (baseline)	0.792	0.777	0.090	0.066
FAN ($\lambda_{\mathrm{EO}} = 0$)	0.916	0.806	0.026	0.004
FAN ($\lambda_{\mathrm{EO}} = 1$)	0.913	0.802	0.038	0.002
FAN ($\lambda_{\mathrm{EO}} = 2$)	0.912	0.797	0.029	0.002
FAN ($\lambda_{\mathrm{EO}} = 3$)	0.911	0.798	0.022	0.002

Table 19: 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 ($a = 0$).

N 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 $\mathrm{TPR}_g = \Pr(\widehat{y} = 1 \mid Y = 1, A = g)$ and $\mathrm{FPR}_g = \Pr(\widehat{y} = 1 \mid Y = 0, A = g)$ for $g \in \{0, 1\}$ (non-minority, minority), together with the gaps $\Delta\mathrm{TPR} = \mathrm{TPR}_1 - \mathrm{TPR}_0$ and $\Delta\mathrm{FPR} = \mathrm{FPR}_1 - \mathrm{FPR}_0$. We summarize the equalized-odds gap for a subgroup as $\max\{|\Delta\mathrm{TPR}|, |\Delta\mathrm{FPR}|\}$ 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 $\Delta\mathrm{TPR}$) are excluded from the corresponding cell. Counts by outcome are shown in the detailed tables.

Table 20 reports the 95th 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 $\lambda{=}0$ to $\lambda{=}2$ yields sizable improvements in the most problematic partitions. The 95th percentile gap for race$\times$purpose falls from 0.520 to 0.378, and for race$\times$loan_type from 0.199 to 0.189 with a further decline to 0.116 at $\lambda{=}3$. Race$\times$sex shows small but consistent reductions (from 0.192 to about 0.186). For MSAs, the 95th percentile decreases from 0.179 at $\lambda{=}0$ to 0.162 at $\lambda{=}2$, with a small uptick to 0.166 at $\lambda{=}3$. 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 21 and 22 provide subgroup-level detail for $\lambda{=}0$ and $\lambda{=}3$. Within the sex partition (Table 21), 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 $\Delta\mathrm{TPR}$ and $\Delta\mathrm{FPR}$. 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 ($\Delta\mathrm{TPR}{=}0.20$ 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 22), the largest gaps at $\lambda{=}0$ 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 20, shows that improvements at the geographic level are also present. The 95th percentile of the equalized-odds gap across the top 25 MSAs declines when moving from $\lambda{=}0$ to $\lambda{=}2$, with a slight reversal at $\lambda{=}3$. 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 21 and 22 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	95th pct (race×sex)	95th pct (race×purpose)	95th pct (race×loan_type)	95th pct (top-25 MSAs)
FAN ($\lambda = 0$)	0.1924	0.5196	0.1988	0.1786
FAN ($\lambda = 1$)	0.1878	0.5133	0.1964	0.1759
FAN ($\lambda = 2$)	0.1853	0.3777	0.1887	0.1618
FAN ($\lambda = 3$)	0.1866	0.3708	0.1164	0.1656

Table 20: 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 21: 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 22: EO gaps within loan purpose subgroups

O ROBUSTNESS: DECOMPOSITION OF THE GLOBAL SCORE GAP

Above, 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

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

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

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

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 16 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 $\lambda{=}0$ 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 $\lambda{=}3$ 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 $\lambda{=}0$, `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 $\lambda{=}3$, 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 $\lambda{=}3$ ($\Delta^{(1)}$: 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.

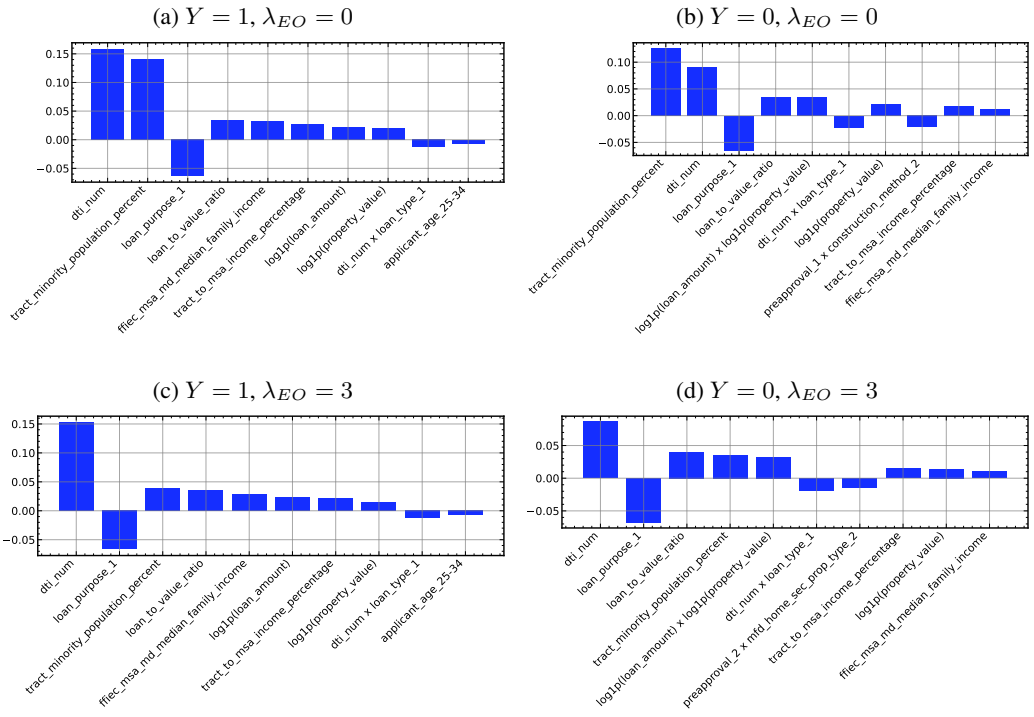


Figure 16: $\Delta_t^{(y)}$ across class labels and models. This figure compares the most important contributors to the score gap for different outcomes and penalization levels. For each case, the ten largest $\Delta_t^{(y)}$ (by absolute value) are reported.