FINDR: TRANSPARENT AND FAIR CREDIT RISK DECISIONS THROUGH SEMI-STRUCTURED REGRESSIONS

A PREPRINT

Victor Medina-Olivares
University of Edinburgh, Edinburgh, UK
victor.medina@ed.ac.uk

Stefan Lessmann
Humboldt-Universität zu Berlin, Berlin, Germany

Jonathan Crook
University of Edinburgh, Edinburgh, UK

ABSTRACT

Credit risk models increasingly need to combine predictive accuracy with transparent explanations and auditable fairness constraints. Logistic regression remains attractive because its coefficients are easy to interpret, but it can miss nonlinear structure. Flexible models can improve prediction, but their explanations are often post-hoc and may not describe the decision rule itself. We introduce `findr`, short for flexible, interpretable deep regression, a semi-structured framework for binary credit risk modelling that decomposes the logit into an interpretable structured component and an orthogonal neural residual. The orthogonalisation separates coefficient-based effects from residual nonlinear variation, while an in-processing Wasserstein penalty mitigates group disparities by comparing score distributions during training. The framework also includes diagnostics that measure the structured component's contribution to logit variation, decision agreement, and local directional consistency. We evaluate `findr` in a simulation study and on eight public credit datasets using score-level accuracy-fairness frontiers. The results show that `findr` behaves close to logistic regression when the signal is approximately linear, while recovering much of the predictive gain of neural models when nonlinear structure is relevant. The diagnostics identify when coefficient-based explanations remain close to the full fitted model and when residual variation must also be examined. These findings support semi-structured modelling as a practical way to make performance, fairness, and interpretability trade-offs explicit in credit risk decisions.

Keywords Semi-structured regression · Deep Neural Networks · Fairness · Interpretability · Credit risk management

1 Introduction

Standard logistic regression remains the workhorse of credit risk modelling, valued for its interpretability, auditability, and well-established regulatory standing (Dumitrescu et al., 2022; Dastile et al., 2020). In its strictly linear log-odds specification, however, it tends to underfit the nonlinearities and feature interactions present in credit portfolios (Dakovic et al., 2010; Djeundje and Crook, 2019; Medina-Olivares et al., 2024). Generalised additive models (GAMs) extend the same exponential-family regression logic, using additive smooth functions rather than only linear covariate effects (Hastie and Tibshirani, 1990). In the binary case, this can be viewed as relaxing the linear predictor of logistic regression while retaining the logistic link and a degree of functional interpretability, but GAMs do not naturally capture higher-order feature interactions unless these are explicitly specified. Moreover, neither GAMs nor standard logistic regression are naturally suited to incorporating the rich unstructured data sources now available to lenders, such as transactional history, text, and images (Stevenson et al., 2021; Korangi et al., 2023). Modern machine learning models address these representational limitations and often deliver strong predictive performance (Baesens et al., 2026), but at the cost of reduced transparency, since they lack inherently auditable decision logic and typically depend on post-hoc tools for interpretation (Gramegna and Giudici, 2021).

This move from transparent statistical models to more flexible machine-learning systems has also changed the governance problem faced by lenders. Recent AI governance initiatives, banking supervision, and fair-lending frameworks increasingly emphasise explainability, auditability, model-risk control, and the monitoring of discriminatory or disparate impacts (European Parliament and Council of the European Union, 2024; European Banking Authority, 2020; Consumer Financial Protection Bureau, 2026). While these sources generally do not prescribe a unique mathematical fairness criterion, they reinforce the need for credit-risk models whose predictive gains can be reconciled with transparent and accountable disparity mitigation.

In the algorithmic fairness literature, these concerns are commonly formalised through criteria that compare outcomes, errors, or score distributions across protected groups[1] (Hurlin et al., 2026; Kim et al., 2023). A highly accurate model may still produce unfair outcomes for protected groups. Therefore, it is crucial to design systems that are both accurate and fair. Bias mitigation approaches are commonly grouped into *pre-processing*, *in-processing*, and *post-processing*. Pre-processing addresses discriminatory patterns before training (Chen et al., 2023). In-processing optimises a fairness objective during training (Wan et al., 2023). Post-processing adjusts model outputs after training (Lohia et al., 2019). In this work, we focus on in-processing fairness, which directly aligns the training objective with fairness goals.

There are many definitions of fairness (Mitchell et al., 2021), which can be broadly categorised into *individual* and *group* notions. **Individual fairness** requires that *similar individuals receive similar outcomes*. It offers person-level protection and helps avoid harms that averages can hide. However, it requires a well-justified similarity metric, and defining and enforcing that metric is domain- and value-dependent and can be computationally costly at scale. **Group fairness** is often more directly auditable at the portfolio level, since firms can compare selection rates, error rates, or score distributions across protected groups. However, group-level summaries can miss within-group disparities ("fairness gerrymandering"), and several group criteria cannot generally be satisfied simultaneously when outcome prevalence differs across groups and predictions are imperfect (Barocas et al., 2023; Chouldechova, 2017). While both paradigms have merits and limitations, we focus on *group-level fairness* as our primary target, given its relevance for portfolio-level model governance and the monitoring of group-level disparities.

Following Barocas et al. (2023), most group fairness criteria fall into three categories, namely *independence*, *separation*, and *sufficiency*. **Independence** (demographic parity) asks that predictions do not differ systematically across protected groups, for example, that approval rates are similar regardless of group membership. **Separation** (error-rate parity) asks that, once the true outcome is fixed, the model behaves similarly across groups. Common special cases are *equalised odds*, which matches both false-positive and true-positive rates, and *equal opportunity*, which matches true-positive rates only. **Sufficiency** (calibration within groups) asks that a given risk score carries the same meaning in every group, so that applicants with a score of 0.2 default at roughly 20% regardless of group membership. This taxonomy clarifies important tensions. When outcome prevalence differs across groups and predictions are imperfect, separation-type guarantees generally conflict with sufficiency, and independence can conflict with either. We therefore treat these criteria as complementary lenses rather than a single target. Our approach accommodates independence, equal opportunity, and equalised odds by selecting the relevant group-conditional score distributions. While sufficiency is valuable for the interpretability of risk scores, it often serves poorly as a standalone fairness target and need not be directly optimised via in-processing penalties (Barocas et al., 2023; Kozodoi et al., 2022).

We propose `findr`, a framework that integrates interpretability, flexibility, and fairness mitigation in a single end-to-end model. The name abbreviates flexible, interpretable deep regression and reflects the aim of combining deep residual flexibility with coefficient-based interpretation. The first contribution is the model itself. It decomposes the log-odds into a structured component, defined by standard interpretable features, and an unstructured component represented by a flexible representation, in our case, a neural network. Orthogonality between the two components is enforced to guarantee identifiability and transparent attribution of effects (Rügamer et al., 2024). Fairness is encoded by penalising Wasserstein distances between the score distributions of sensitive groups, following Wasserstein fair classification (Jiang et al., 2020). A second contribution is evaluative. We introduce interpretability diagnostics that quantify the structured component's contribution to score variability and assess whether structured-component interpretations remain reliable for decisions and feature effects. We also construct score-level accuracy–fairness frontiers that provide an empirical benchmark for comparing fitted model paths at the same Wasserstein fairness level. The framework is evaluated in a simulation study and applied to eight publicly available credit-default datasets.

[1]Throughout the paper, *fairness* refers to formal algorithmic criteria used to assess whether model scores, decisions, or errors differ systematically across protected groups. We distinguish this from *equity*, which is a broader normative and legal concern about whether opportunities, burdens, and remedies are distributed appropriately in context. Our focus on fairness metrics is therefore an operational modelling choice. These metrics capture specific, auditable aspects of equity-relevant disparity, but they do not by themselves establish that a lending system is equitable.

The following section reviews related literature. Section three describes the model, the interpretability diagnostics, and the accuracy–fairness frontier. Section four evaluates the framework using simulated data, and section five applies it to eight real obligor-level datasets. Section six concludes.

2 Related Work

We organise the review around the two areas that motivate the framework. We first discuss interpretable credit-risk models and the question of how to explain flexible predictors. We then discuss statistical fairness criteria and penalties that reduce score-level disparities.

Credit scoring has long relied on logistic regression because it is probabilistic, auditable, and easy to communicate (Lessmann et al., 2015; Dastile et al., 2020). This transparency comes from a restrictive linear form for the log-odds, which can miss nonlinearities and interactions in credit portfolios (Dakovic et al., 2010; Dumitrescu et al., 2022). More flexible models can improve predictive fit, but their explanations often rely on tools applied after the model is fitted (Gramegna and Giudici, 2021). Methods based on rule extraction, distillation, and surrogate models can approximate complex models with more interpretable structures (Martens et al., 2007; Tan et al., 2023; Frosst and Hinton, 2017). Their limitation is that the explanation is not the fitted function itself. It can be informative at a global level while remaining imperfect for the local decisions that matter in lending.

Models that are interpretable by design respond to this concern by making the explanation part of the predictor itself, rather than an external approximation fitted after the fact. Generalised additive models provide readable feature effects and can relax linearity while still writing the prediction as a sum of feature effects (Hastie and Tibshirani, 1990; Djeundje and Crook, 2019). GA2M-style models and neural additive models extend this idea through pairwise interactions or neural feature functions (Lou et al., 2013; Agarwal et al., 2021). GAMI-Net and NODE-GAM further develop this line by adding structured interactions or additive architectures that can be trained by gradient methods (Yang et al., 2021; Chang et al., 2022). These models weaken the simple accuracy-interpretability trade-off, but their transparency still depends on restrictions on how effects and interactions enter the predictor. Semi-structured regression takes a different route by combining a conventional structured predictor with a flexible neural component (Rügamer et al., 2024). This architecture is attractive for credit risk because the structured term can retain coefficient effects, while the neural residual captures variation that is not well represented by the structured specification. Related semi-structured ideas have been used for time to default with cure fraction and multi-state delinquency modelling (Medina-Olivares et al., 2024, 2026).

The gain in flexibility creates a problem of attribution. If the structured and neural components can represent the same signal, an apparently interpretable coefficient may depend on an arbitrary allocation of fitted variation rather than on a property of the overall predictor. Orthogonalisation addresses this overlap by removing from the neural representation the part that can be explained by the structured variables (Rügamer et al., 2024). This gives a well-defined separation between effects assigned to the coefficients and residual patterns assigned to the neural component. This differs from approaches that make additive feature effects simpler or more structured (Luber et al., 2023). Their target is the form of the feature functions, while our target is the share of the integrated predictor that remains attributable to the structured component. Yet, identifiability is only a necessary condition for interpretation. Orthogonality does not show that the structured specification captures the relevant structure in the data, that coefficients are stable under distribution shift, or that the neural residual is small. A model can have uniquely defined coefficients while most score variation or some decisions are driven by the residual. We therefore treat interpretability as an empirical property of the fitted decomposition, assessed by how much logit variation comes from the structured component, how often structured and full-model decisions agree, and whether local effects point in the same direction.

Fairness adds a second challenge because accurate and interpretable models can still produce scores or decisions that differ systematically across protected groups. Bias-mitigation methods intervene before, during, or after model fitting (Hort et al., 2024). Methods that act during training impose fairness through regularisation, constrained optimisation, or adversarial learning (Kamishima et al., 2012; Zafar et al., 2017; Agarwal et al., 2018; Zhang et al., 2018). Methods applied after training can satisfy a criterion at a fixed threshold, but lending thresholds may vary across products, portfolios, economic conditions, or resource constraints (De Vos et al., 2025). Training-time methods can instead act directly on the probabilistic score. This distinction matters because fairness at one decision threshold does not imply similarity of the underlying score distributions.

The choice of fairness criterion is also a subject of debate. Independence, separation, and sufficiency encode different requirements and generally cannot all hold when group base rates differ and prediction is imperfect (Hardt et al., 2016; Chouldechova, 2017; Barocas et al., 2023). Credit-scoring studies show that conclusions about fairness, performance, and profitability depend on both the selected criterion and the mitigation method (Kozodoi et al., 2022; Hurlin et al., 2026). Fairness assessments based on a single protected attribute can also miss intersectional disparities (Kim et al.,

2023). These findings suggest that there is no single criterion-free notion of a fair credit score. Our focus on score distributions conditional on the outcome is therefore a modelling choice, motivated by the role of error behaviour in credit-risk decisions.

Many fairness penalties that can be optimised by gradient methods compare group means or smooth approximations to threshold-based errors. These objectives are tractable, but they do not control the full distribution of scores. Two groups can have equal average predictions while differing in dispersion, tails, or the share of cases that would cross alternative decision thresholds. Fair Mixup addresses the generalisation of fairness constraints by regularising expected predictions along interpolation paths between group distributions (Chuang and Mroueh, 2021). Wasserstein fair classification compares group score distributions more directly (Jiang et al., 2020). For univariate risk scores, the Wasserstein distance has an exact quantile representation and captures discrepancies throughout the score distribution (Villani et al., 2008; Peyré and Cuturi, 2019; Panaretos and Zemel, 2019).

This use of optimal transport differs from distributionally robust fairness, where robustness refers to performance under plausible shifts in the data distribution. Wasserstein distributionally robust methods define neighbourhoods around the empirical data distribution and seek models that retain performance or fairness under adverse perturbations (Taskesen et al., 2020; Wang et al., 2020). Our penalty uses Wasserstein distance for a different object. It compares protected groups within the fitted score space. The former addresses uncertainty about the population distribution, while the latter directly targets between-group disparity in model outputs.

These strands leave a gap at the intersection of fairness, flexibility, and attribution. Explanations fitted after training can describe a black box without being the decision rule. Intrinsic additive models make interpretation part of the predictor but restrict the remaining interaction structure. Semi-structured models restore flexibility and identify how fitted signal is allocated between structured and neural components, but identifiability alone does not show that the structured term explains the final score or decision. Fairness-aware classifiers can reduce selected disparities, but interventions at a fixed threshold do not control the score distribution and usually do not ask which component of a hybrid predictor absorbs the fairness adjustment.

`findr` addresses these issues jointly. It combines an identifiable structured logit with an orthogonal neural residual and applies an in-processing Wasserstein penalty to the complete probability score. Fairness is imposed on the predictor that is ultimately deployed, rather than on the structured component alone or on decisions at one threshold. At the same time, diagnostics assess whether the resulting fair score remains meaningfully attributable to its coefficient-based component. The score-level frontier then places fitted model paths on a common empirical accuracy–fairness scale. The contribution is to examine whether nonlinear fairness mitigation and structured explanation can coexist, and to make the trade-off between distributional fairness, predictive performance, and the completeness of coefficient-based attribution visible.

3 Methodology

3.1 Semi-structured logistic model

For observation $i = 1, \ldots, n$, let $y_i \in \{0, 1\}$ denote the default indicator (1 = default) and let $A_i \in \{0, 1\}$ denote a sensitive attribute (1 = sensitive). We develop the method for this binary sensitive-attribute case, while the same construction can be extended to multi-group sensitive attributes. The vector $\boldsymbol{x}_i \in \mathbb{R}^p$ contains the covariates used in the structured linear component. The vector $\boldsymbol{z}_i$ contains the inputs to the neural network, which may coincide with $\boldsymbol{x}_i$, add further variables, or represent a separate source of information (e.g., text data). Following the semi-structured regression formulation of Rügamer et al. (2024), we decompose the total log-odds η_i into a structured contribution η_i^{str} and an unstructured contribution η_i^{unstr}:

$$\eta_i \;=\; \eta_i^{\mathrm{str}} \;+\; \eta_i^{\mathrm{unstr}}, \qquad \eta_i^{\mathrm{str}} \;=\; \beta_0 + \boldsymbol{x}_i^{\top}\boldsymbol{\beta}, \qquad \eta_i^{\mathrm{unstr}} \;=\; d_\theta(\boldsymbol{z}_i), \tag{1}$$

where η_i^{str} is a linear logit with explicit coefficient effects, while η_i^{unstr} is a neural-network residual on the same logit scale. Here, d_θ denotes the full unstructured logit map. In the identifiable version below, this map is implemented by first applying an encoder h_α, then orthogonalising its sample-level output, and finally applying a residual layer r_γ, with $\theta = (\alpha, \gamma)$. The default probability is $p(y_i = 1 \mid \eta_i) = \sigma(\eta_i)$, where $\sigma(t) = (1 + \exp(-t))^{-1}$. Extensions to multiple unstructured components and generalisations to models for location, scale, and shape are also possible (Thielmann et al., 2024).

3.2 Identifiability via orthogonalisation

When the same covariates enter both the structured and unstructured components in (1), the decomposition is not identifiable without additional constraints. Linear effects can then be represented either by the coefficient vector $\boldsymbol{\beta}$ or

by the neural network, which makes attribution to the structured component ambiguous. To prevent this, we project the neural representation away from the linear span of the structured design before forming the residual logit.

Let $\mathbf{1}_n \in \mathbb{R}^n$ denote the all-ones vector, and define the structured design matrix with intercept

$$\tilde{\boldsymbol{X}} = [\mathbf{1}_n \;\; \boldsymbol{X}] \in \mathbb{R}^{n\times(p+1)}, \qquad \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}.$$

With $\tilde{\boldsymbol{\beta}} = (\beta_0, \boldsymbol{\beta}^\top)^\top$, the structured logit vector is $\boldsymbol{\eta}^{\mathrm{str}} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}}$. Assume $\tilde{\boldsymbol{X}}$ has full column rank and $n \geq p+1$. Take the thin QR decomposition

$$\tilde{\boldsymbol{X}} = \boldsymbol{Q}\boldsymbol{R}, \qquad \boldsymbol{Q} \in \mathbb{R}^{n\times(p+1)} \text{ with } \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_{p+1}, \qquad \boldsymbol{R} \in \mathbb{R}^{(p+1)\times(p+1)} \text{ upper triangular.}$$

The orthogonal projector onto the column space of $\tilde{\boldsymbol{X}}$ is

$$\mathcal{P} = \boldsymbol{Q}\boldsymbol{Q}^\top \in \mathbb{R}^{n\times n}, \qquad \mathcal{P}^\perp = \boldsymbol{I}_n - \mathcal{P} \in \mathbb{R}^{n\times n}.$$

The matrix $\mathcal{P}^\perp$ maps any sample-level vector onto the orthogonal complement of the column space of $\tilde{\boldsymbol{X}}$. Writing $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$, let $\boldsymbol{H}_\alpha = (h_\alpha(\boldsymbol{z}_1), \ldots, h_\alpha(\boldsymbol{z}_n))^\top \in \mathbb{R}^{n\times q}$ collect the neural encoder outputs before the final residual layer. Orthogonalisation is applied columnwise as

$$\boldsymbol{H}_\alpha^\perp = \mathcal{P}^\perp \boldsymbol{H}_\alpha. \tag{2}$$

Let r_γ denote the final residual layer, with coefficients $\boldsymbol{\gamma} \in \mathbb{R}^q$ and intercept γ_0. In our implementation this layer is linear, so the unstructured contribution vector is

$$\boldsymbol{\eta}^{\mathrm{unstr}} = d_\theta(\boldsymbol{Z}) = r_\gamma(\boldsymbol{H}_\alpha^\perp) = \boldsymbol{H}_\alpha^\perp\boldsymbol{\gamma} + \gamma_0\mathbf{1}_n. \tag{3}$$

Here $d_\theta(\boldsymbol{Z})$ denotes the row-wise application of the full unstructured map to the sample. Since $\tilde{\boldsymbol{X}}^\top\boldsymbol{H}_\alpha^\perp = \mathbf{0}$, the projected encoder features contain no intercept term and no linear combination of the covariates in $\boldsymbol{X}$. Apart from the residual-layer intercept γ_0, which can be absorbed into the structured intercept, linear effects are therefore attributed to the structured component, and the neural residual captures variation outside the structured linear subspace.

3.3 In-processing fairness via Wasserstein distances

Post-hoc fairness corrections, such as calibrating group-specific thresholds after training, can satisfy a criterion at a chosen deployment threshold but need not remain valid when thresholds change (Jiang et al., 2020). This is important in credit scoring, where thresholds shift with loan capacity, risk appetite, and regulatory requirements (De Vos et al., 2025). We therefore impose the fairness penalty on the score distributions themselves. This does not guarantee a specific value of the chosen fairness criterion for every subsequent threshold, but it reduces reliance on threshold-specific post-processing by acting on the distributions from which thresholded decisions are derived.

Let $\widehat{Y}_i = \sigma(\hat{\eta}_i) \in [0,1]$ denote the predicted probability of default for observation i, where $\hat{\eta}_i$ is the fitted logit. Group fairness criteria compare the distributions of $\widehat{Y}$ across groups defined by A, optionally conditioning on the true outcome Y (Barocas et al., 2023). We quantify between-group differences using the p-Wasserstein distance on $\mathbb{R}$, where the order p is usually taken to be 1 or 2 (Shalit et al., 2017).[2]

We first define the empirical Wasserstein distance used in the penalty. Consider two empirical probability measures on $\mathbb{R}$,

$$\hat{\mu} = \sum_{i=1}^{n} \omega_i\, \delta_{s_{(i)}}, \qquad \hat{\nu} = \sum_{j=1}^{m} \upsilon_j\, \delta_{t_{(j)}},$$

with sorted supports $s_{(1)} \leq \cdots \leq s_{(n)}$, $t_{(1)} \leq \cdots \leq t_{(m)}$ and weights $\omega \in \Delta^n$, $\upsilon \in \Delta^m$. For a ground metric $d(s,t) = |s-t|$, the discrete p-Wasserstein problem is

$$W_p^p(\hat{\mu}, \hat{\nu}) \;=\; \min_{T\in\Pi(\omega,\upsilon)} \sum_{i,j} T_{ij}\, |s_{(i)} - t_{(j)}|^p, \qquad \Pi(\omega,\upsilon) = \{T \geq 0 :\; T\mathbf{1} = \omega,\; T^\top\mathbf{1} = \upsilon\}, \tag{4}$$

which equals the minimum average cost of transporting $\hat{\mu}$ into $\hat{\nu}$ (Villani et al., 2008).

Because credit scores are one-dimensional, W_p admits a closed-form characterisation via quantile matching (Peyré and Cuturi, 2019) and no entropic regularisation is required. Let $\Omega_i = \sum_{r\leq i} \omega_r$ and $\Upsilon_j = \sum_{\ell\leq j} \upsilon_\ell$ denote the cumulative

[2] Our simulations and empirical analysis focus on $p = 1$, but the same construction can be directly used with other Wasserstein orders.

weights of the two empirical distributions. Let $0 = q_0 < q_1 < \cdots < q_K = 1$ be the ordered grid of cumulative-mass breakpoints obtained from $\{\Omega_i\}_{i=1}^n$ and $\{\Upsilon_j\}_{j=1}^m$, with 0 added as the left endpoint. Writing the empirical quantiles

$$F_{\hat{\mu}}^{-1}(q) = s_{(i)} \text{ for } i = \min\{r : \Omega_r \geq q\}, \qquad F_{\hat{\nu}}^{-1}(q) = t_{(j)} \text{ for } j = \min\{\ell : \Upsilon_\ell \geq q\},$$

we obtain the exact discrete formula

$$W_p^p(\hat{\mu}, \hat{\nu}) \;=\; \sum_{k=1}^{K} (q_k - q_{k-1}) \left| F_{\hat{\mu}}^{-1}(q_{k-1}) - F_{\hat{\nu}}^{-1}(q_{k-1}) \right|^p. \tag{5}$$

This one-dimensional form is useful for our training objective because it evaluates the fairness penalty exactly by sorting the scores and is differentiable almost everywhere, allowing gradients to be propagated through the Wasserstein term. In what follows, we write W for the chosen p-Wasserstein distance, with $p = 1$ in our empirical analysis.

We can now express the group fairness criteria from above as score-level penalties. The choice of criterion determines which group-conditional distributions of $\widehat{Y}$ are compared.

Equalised Odds: $\displaystyle\sum_{y \in \{0,1\}} W\big(\, \widehat{Y} \mid Y{=}y, A{=}0 \;,\; \widehat{Y} \mid Y{=}y, A{=}1 \,\big),$

Equal Opportunity: $W\big(\, \widehat{Y} \mid Y{=}0, A{=}0 \;,\; \widehat{Y} \mid Y{=}0, A{=}1 \,\big),$

Independence: $W\big(\, \widehat{Y} \mid A{=}0 \;,\; \widehat{Y} \mid A{=}1 \,\big).$ (6)

Equalised odds compares score distributions separately within each realised outcome group, so disparities are penalised among both non-defaulters and defaulters. Equal opportunity focuses only on the outcome group treated as opportunity-relevant in this application, here $Y = 0$, and therefore targets score disparities among non-defaulters. Independence does not condition on the outcome and instead compares the overall score distributions across protected groups.

For a chosen criterion, we set $\mathcal{L}_{\text{fair}}(\widehat{Y}, A, Y)$ to the corresponding scalar penalty in (6), computed via (5).

3.4 Training objective

The total loss is a weighted combination of predictive fit and a fairness penalty,

$$\mathcal{L}_{\text{total}}(\theta, \beta) \;=\; \underbrace{\frac{1}{n} \sum_{i=1}^{n} \big[-y_i \log \widehat{Y}_i - (1 - y_i) \log(1 - \widehat{Y}_i) \big]}_{\mathcal{L}_{\text{pred}}} \;+\; \lambda \cdot \mathcal{L}_{\text{fair}}\big(\widehat{Y}, A, Y\big), \tag{7}$$

with $\lambda \geq 0$ governing the accuracy-fairness trade-off. Gradients are backpropagated through both the orthogonalisation step and the Wasserstein distance to update (θ, β) via stochastic optimisation. Figure 1 summarises the resulting architecture.

3.5 Score-level accuracy–fairness frontier

The training objective above produces a family of fitted models indexed by the fairness multiplier λ. To evaluate these models, we compare them with a score-fairness reference frontier. The frontier asks how much predictive accuracy is attainable at a given value of the same Wasserstein unfairness criterion used for training. This differs from standard accuracy–fairness frontiers, which are usually defined for thresholded decisions through false-positive and false-negative rates (Hardt et al., 2016; Menon and Williamson, 2018). Our penalty is applied before thresholding, so the fairness axis compares distributions of probability scores.

Let $\mathcal{D}_{\text{val}} = \{(x_i, y_i, a_i)\}_{i=1}^n$ be the validation sample. Let $r_i \in (0, 1)$ denote a reference estimate of $P(Y = 1 \mid X = x_i)$, using the true conditional risk in simulations when available and otherwise the score from an unconstrained model fitted on the training data. A candidate predictor assigns one score $s_i \in (0, 1)$ to each validation observation. This mirrors how fitted models are evaluated, since each individual receives one probability score, and the realised labels and groups are then used only to compute prediction loss and fairness gaps.

For equalised odds, the empirical unfairness of a candidate score is

$$\widehat{\Phi}(s; y, a) = \sum_{y \in \{0,1\}} W\big(\widehat{\mu}_{y0}^s, \widehat{\mu}_{y1}^s\big)\,, \tag{8}$$

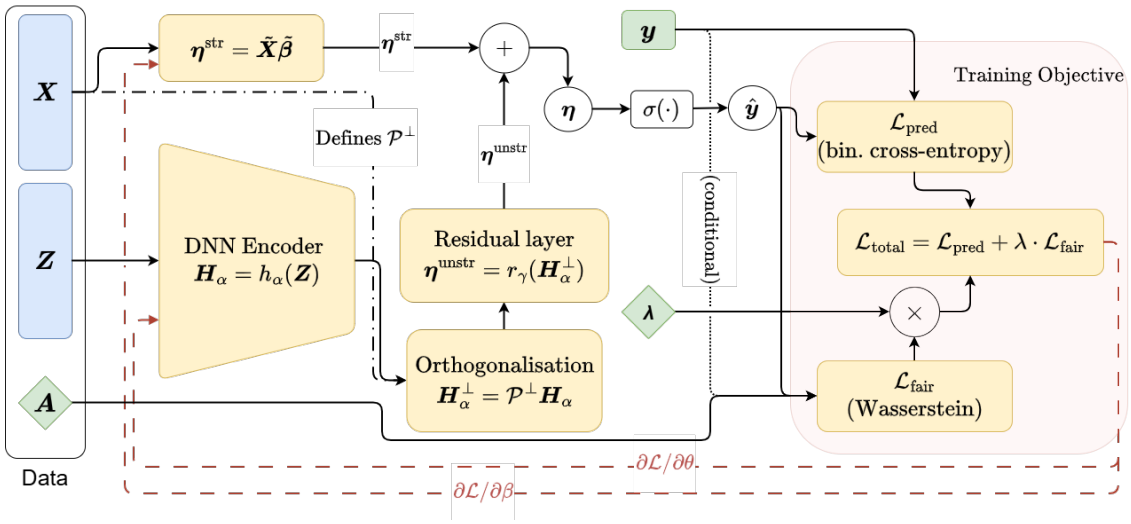

Figure 1: `findr` architecture.

where $\widehat{\mu}^{s}_{ya}$ is the empirical distribution of s_i among validation observations with $Y_i = y$ and $A_i = a$. For $p = 1$, this equals the integral of the absolute difference between the two empirical CDFs over all score cutoffs. Thus y_i and a_i enter the frontier through the fairness functional, because equalised odds is defined by distributions conditional on realised outcome and sensitive group. We focus on equalised odds here, but the same procedure applies to the other criteria in (6) by replacing $\widehat{\Phi}$ with the corresponding empirical score-distribution penalty.

Predictive accuracy is measured by the expected Bernoulli log loss of s under the reference risk r,

$$\widehat{R}_{\log}(r, s) = -\frac{1}{n}\sum_{i=1}^{n}\big[r_i \log s_i + (1 - r_i)\log(1 - s_i)\big] = \widehat{H}(r) + \frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\{\mathrm{Bern}(r_i) \,\|\, \mathrm{Bern}(s_i)\}. \tag{9}$$

Here $\widehat{H}(r)$ is the empirical Bernoulli entropy of the reference risks, which is the expected log loss obtained by predicting r_i itself. The second equality therefore writes the expected log loss as this irreducible term plus the average KL divergence from r to s. This expression does not use the realised y_i because it is the conditional expectation of binary log loss given X_i under $P(Y = 1 \mid X_i) = r_i$. It therefore measures the expected loss relative to the reference risk surface rather than sampling noise in a particular validation draw. The unconstrained optimum is $s = r$, and plots report the excess log loss $\widehat{R}_{\log}(r, s) - \widehat{H}(r)$ so that this point has value zero.

The predictor frontier is the lower boundary of

$$\min_{s} \widehat{R}_{\log}(r, s) \quad \text{subject to} \quad \widehat{\Phi}(s; y, a) \le \tau,$$

as τ varies. Numerically, we trace this boundary by solving the scalarised problem

$$\min_{s} \; Q_\xi(s) = \widehat{R}_{\log}(r, s) + \xi\, \widehat{\Phi}(s; y, a), \qquad \xi \ge 0, \tag{10}$$

over a grid of frontier multipliers ξ. We use ξ rather than λ to separate the post-hoc frontier construction from the model-training multiplier in (7). At $\xi = 0$, the solution is $s = r$. Larger values of ξ move the score vector toward lower empirical unfairness, usually at the cost of higher expected log loss. The resulting points, together with any candidate model scores, are sorted by $\widehat{\Phi}$ and converted to a monotone lower envelope. Appendix A gives the computational details. The frontier should be interpreted as a numerical score-vector benchmark rather than as the attainable boundary for any specified family of fitted models. It gives the best empirical trade-off found by the construction on the validation sample, relative to the reference risk r, and is therefore useful for comparing fitted models on a common scale rather than for establishing a population-optimal accuracy–fairness bound.

For a fitted model with validation scores s_m, the accuracy gap is

$$\Delta(s_m) = \widehat{R}_{\log}(r, s_m) - F\Big(\widehat{\Phi}(s_m; y, a)\Big), \tag{11}$$

where F is the interpolated frontier value. A smaller gap means that the fitted model is closer to the best score vector found at the same fairness level. The quantity is empirical and depends on the numerical frontier construction, so small negative values can occur if the computed envelope fails to dominate an unseeded candidate score.

3.6 Diagnostics for interpretable-component contribution

Beyond coefficient readability, we use three diagnostics to quantify the contribution of the interpretable structured component to the fitted model. These diagnostics assess its role in score variation, binary decisions, and feature-level directional interpretations. Throughout this subsection, $\boldsymbol{\eta}$ denotes the full logit vector, $\boldsymbol{\eta}^{\mathrm{str}}$ denotes the structured logit vector, and $\boldsymbol{\eta}^{\mathrm{unstr}}$ denotes the neural residual from (1).

3.6.1 Explained Variability Ratio (EVR)

EVR measures the share of logit-scale variation attributable to the structured component. We define

$$\mathrm{EVR} \;=\; \frac{\mathrm{Var}_n\big(\boldsymbol{\eta}^{\mathrm{str}}\big)}{\mathrm{Var}_n\big(\boldsymbol{\eta}\big)} \;=\; \frac{\frac{1}{n}\sum_i \big(\eta_i^{\mathrm{str}} - \bar{\eta}^{\mathrm{str}}\big)^2}{\frac{1}{n}\sum_i \big(\eta_i - \bar{\eta}\big)^2}. \tag{12}$$

The orthogonalisation in (2) makes the total logit variance decompose into the sum of structured and residual variation. Hence

$$\mathrm{EVR} = \frac{\mathrm{Var}_n(\boldsymbol{\eta}^{\mathrm{str}})}{\mathrm{Var}_n(\boldsymbol{\eta}^{\mathrm{str}}) + \mathrm{Var}_n(\boldsymbol{\eta}^{\mathrm{unstr}})}.$$

Thus EVR can be interpreted as the structured component's share of total logit-scale variation. A high EVR indicates that most score variation is explained by the transparent linear component. A low EVR indicates that the neural residual accounts for a material share of the fitted score variation.

3.6.2 Decision Disagreement Rate (DDR)

EVR is measured on the continuous logit scale. DDR complements it by measuring whether the structured component and the full model imply the same binary decision at the default threshold. Let

$$\hat{y}_i \;=\; \mathbb{1}[\eta_i \geq 0], \qquad \hat{y}_i^{\mathrm{str}} \;=\; \mathbb{1}[\eta_i^{\mathrm{str}} \geq 0]$$

denote the class labels implied by the full and structured logits. The decision disagreement rate is

$$\mathrm{DDR} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(\hat{y}_i^{\mathrm{str}} \neq \hat{y}_i\big)\,. \tag{13}$$

DDR is zero when the structured component and the full model make identical classifications on the evaluation sample. Larger values indicate that using the structured component alone would often lead to different decisions than using the integrated model.

3.6.3 Counterfactual Sign Error Rate (CSER)

The structured coefficients also provide feature-level directional interpretations. CSER checks whether these directions remain consistent after adding the neural residual. For each structured feature j with $\beta_j \neq 0$, let $d_j = \mathrm{sign}(\beta_j)$ and let $\delta > 0$ denote a prespecified perturbation size. Define

$$\boldsymbol{x}_i^{(j,+)} = \boldsymbol{x}_i + d_j\,\delta\,\boldsymbol{e}_j,$$

where $\boldsymbol{e}_j$ is the jth unit vector. This perturbation moves feature j in the direction that the structured coefficient associates with higher default log-odds. The counterfactual sign error rate for feature j is

$$\mathrm{CSER}_j \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\,\eta(\boldsymbol{x}_i^{(j,+)},\boldsymbol{z}_i) < \eta(\boldsymbol{x}_i,\boldsymbol{z}_i)\,\}}. \tag{14}$$

CSER_j is the share of observations for which a risk-increasing perturbation according to the structured coefficient lowers the full model logit. It therefore measures directional contradictions between the transparent coefficient and the integrated model. In the tables, we report the minimum and maximum CSER across structured features,

$$\mathrm{CSER}_{\min} = \min_{j:\beta_j\neq 0} \mathrm{CSER}_j, \qquad \mathrm{CSER}_{\max} = \max_{j:\beta_j\neq 0} \mathrm{CSER}_j,$$

to summarise the range of feature-level sign reliability.

4 Simulation Study

The simulation is designed to study three questions. First, we ask whether findr behaves like a transparent linear model when the true signal is linear. Second, we ask whether the neural residual recovers additional predictive structure when the signal is nonlinear. Third, we analyse how the Wasserstein penalty moves fitted scores along the accuracy–fairness trade-off while preserving the contribution of the structured component.

4.1 Data-generating processes

We simulate $n = 10{,}000$ observations with $p = 6$ covariates. The sensitive attribute and pre-shift covariates follow

$$A \sim \mathrm{Bern}(0.3), \qquad X_j \sim \mathcal{N}(0,1) \;\; (j = 1, \ldots, 6).$$

We induce two sources of score-distribution unfairness. The first is a group-dependent covariate shift, $X_1 \leftarrow X_1 + (1 - A)$. The second is a direct group-specific logit shift, $B(A) = 0.2\,\mathbf{1}\{A = 1\}$. The linear component is $L = b_0 + \boldsymbol{\beta}^\top X$, with $b_0 = -2$ and $\boldsymbol{\beta} = (-1, -0.5, 0, 0.25, 0.5, 1)^\top$. To create nonlinear structure, we add univariate nonlinear terms $\mathrm{NL}(X) = \sum_{k \in \mathcal{I}_{\mathrm{nl}}} g_k(X_k)$ and interactions $\mathrm{I}(X) = \sum_{(i,j) \in \mathcal{I}_{\mathrm{int}}} X_i X_j$. We consider two data-generating processes:

- **Simple mode**. $\eta = L + B(A)$, so the true logit is linear with both covariate-shift and baseline unfairness.
- **Complex mode**. $\eta = L + \mathrm{NL}(X) + \mathrm{I}(X) + B(A)$, so the true logit contains linear, nonlinear, and interaction effects, again with both sources of unfairness.

Outcomes are drawn from $Y \sim \mathrm{Bern}(\sigma(\eta))$. The simple mode therefore tests whether the semi-structured model avoids using the residual when it is unnecessary. The complex mode tests whether the same architecture can recover nonlinear signal without losing the linear component used for interpretation.

4.2 Models and training

We compare three models. LR is a standard logistic regression and represents the transparent linear baseline. NN is a two-layer feed-forward neural logistic model with the same neural backbone as findr, but without a separate structured component. findr is the semi-structured model in Section 3, with a linear structured logit and an orthogonal neural residual.

All models are trained with the objective in (7). We fit the three fairness penalties described in (6) in turn. We vary the multiplier λ over a grid from 0 to 4 and repeat each configuration over 10 random seeds. Some visual summaries display only the range up to $\lambda = 1$ when larger penalties do not add substantive information. We report the equalised-odds results because this criterion conditions on both outcome classes, making it the most demanding of the three criteria considered here. The score-level frontier in Section 3.5 is constructed for the same criterion, and the other criteria led to the same qualitative comparisons across models.

4.3 Evaluation

We evaluate the simulations along the same three dimensions used in the rest of the paper. Predictive performance is measured out of sample using AUC and Brier score. Fairness is measured by the equalised-odds Wasserstein penalty, and the frontier places model paths on the common score-level accuracy–fairness scale developed in Section 3.5. Interpretable-component contribution is assessed using the coefficient paths, EVR, DDR, and CSER from the diagnostics in Section 3.

Out-of-sample performance. Figure 2 compares predictive performance as λ varies. In the simple mode, the true logit is linear, so logistic regression is correctly specified. The neural model and findr therefore have little room to improve predictive accuracy over LR. The relevant check is whether findr preserves this linear behaviour rather than introducing unnecessary nonlinear variation. In the complex mode, the neural residual has a role because the data-generating logit includes nonlinear terms and interactions. The expected pattern is therefore different, since LR is misspecified, while NN and findr can use the neural component to recover additional predictive signal.

In the simple mode, findr behaves like the linear model because the orthogonal residual has little unexplained structure to capture. In the complex mode, findr instead follows the neural model more closely than the linear baseline, indicating that the residual component captures nonlinear signal while the structured component remains available for interpretation. As λ increases, performance can decline because the fitted scores move toward smaller score-distribution gaps.

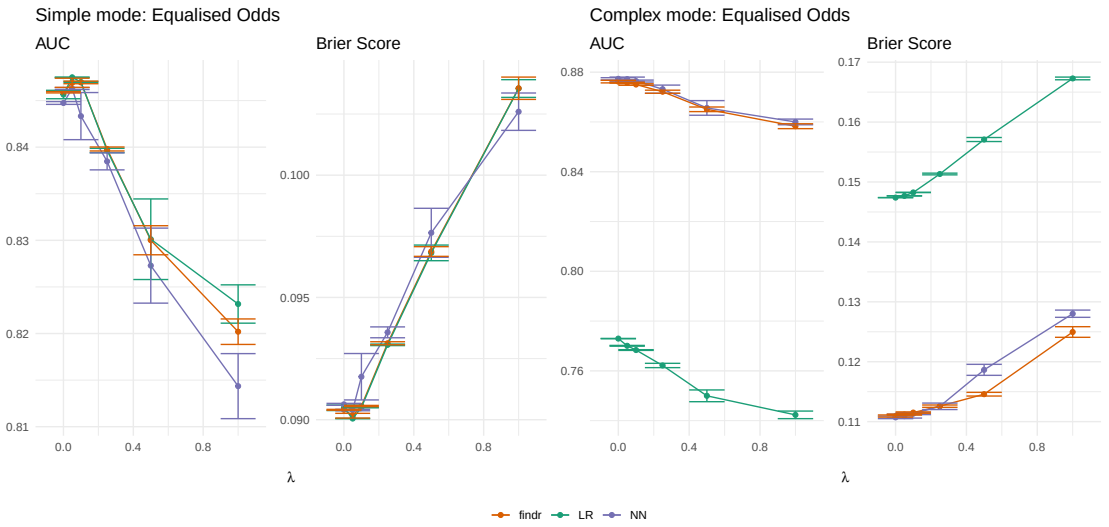

Figure 2: Out-of-sample performance in the simple and complex simulation settings. Curves show the mean across 10 random seeds as the fairness multiplier λ varies under equalised odds.

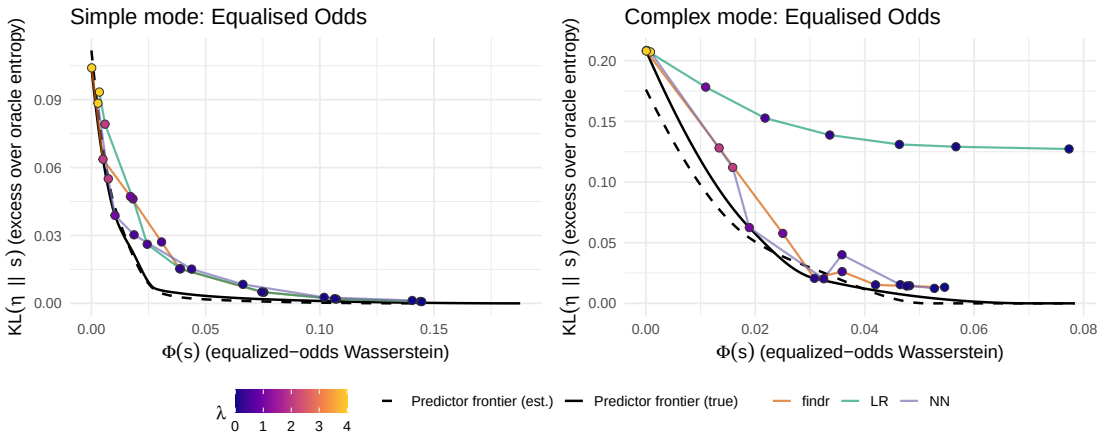

Figure 3: Predictor frontiers and model paths under equalised odds. The panels compare the simple and complex simulation settings. Black curves show the true and estimated predictor frontiers in the fairness–log-loss plane, where the horizontal axis is the equalised-odds Wasserstein distance $\widehat{\Phi}(s; y, a)$ and the vertical axis is the excess log loss relative to the reference-risk entropy $\widehat{H}(r)$. Coloured lines trace the solutions obtained by each model as the regularisation parameter λ varies, and the point colour gradient indicates the value of λ. Points closer to the lower-left region correspond to lower fairness violation and lower predictive loss.

Fairness. Figure 3 places the fitted model paths against the predictor frontier. The horizontal axis is the empirical equalised-odds Wasserstein distance $\widehat{\Phi}(s; y, a)$, and the vertical axis is the excess expected log loss relative to the reference-risk entropy $\widehat{H}(r)$. Points nearer the lower-left region have both smaller fairness violation and smaller prediction loss. The black frontier is a score-vector benchmark rather than a model-class boundary, so the comparison asks how close each fitted model comes to the best score trade-off found on the validation sample.

The model paths illustrate how the in-processing penalty acts on the full score distribution. At $\lambda = 0$, models prioritise predictive fit and can retain the group differences induced by the covariate and baseline shifts. As λ increases, their scores move toward lower equalised-odds distance. The trade-off is mild when the fairness correction is aligned with the available model structure, and stronger when fairness constraints require score changes that also reduce predictive fit.

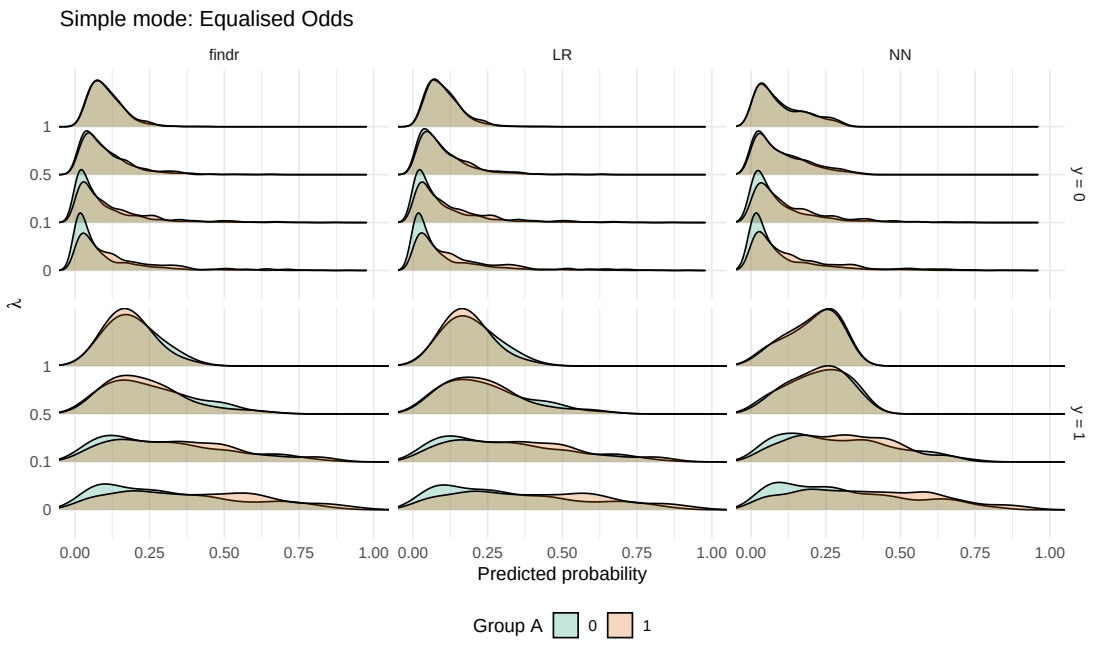

Figure 4: Score distributions in the simple simulation setting. The panels show fitted score distributions across groups and outcome strata as λ varies.

Figures 4 and 5 show the same mechanism at the distributional level. As expected, the penalty does not adjust a single threshold but changes the fitted score distributions within the outcome-by-group cells that define equalised odds. As λ increases, these group-conditional distributions move closer within each model. This does not require the final score distributions to have the same shape across model specifications. Different model classes can satisfy a smaller Wasserstein gap while still producing distinct score distributions. This agrees with the motivation for using Wasserstein penalties in Section 3, where the fairness intervention is applied before thresholding.

Interpretability. We next study how much of the fitted model can be assigned to the structured component. In the simple mode, the true signal is linear, so `findr` should behave like a linear model with little need for the residual. The coefficient paths in Figure 6 follow the same pattern as `LR`, including the signs of the linear effects. The diagnostics in Table 1 lead to the same conclusion. EVR remains at or near one across the reported λ values, DDR is essentially zero, and CSER is negligible. Thus, the full fitted logit is almost entirely explained by the structured component, and using the structured component alone leads to the same decisions for nearly all observations. This is the desired result in a setting where the residual has no genuine nonlinear signal to recover.

Figure 6 also shows how the fairness penalty affects the structured component. Recall that the simulation setup encodes feature-level disparities in covariate X_1. Accordingly, increasing λ shrinks the effect of that feature, β_1, toward zero, whereas the effects of other covariates change relatively little.

Table 1: **Simple mode:** Equalised Odds. EVR, DDR, CSER (min) and CSER (max) reported as mean (SD) for the findr model by λ.

λ	EVR	DDR	CSER (min)	CSER (max)
0.0	1.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0.1	1.000 (0.000)	0.001 (0.001)	0.000 (0.000)	0.000 (0.000)
0.5	0.997 (0.009)	0.002 (0.004)	0.000 (0.000)	0.002 (0.006)
1.0	0.984 (0.019)	0.001 (0.001)	0.000 (0.000)	0.005 (0.014)

In the complex mode, the signal is split by design between linear terms, nonlinear terms, and interactions. The residual should therefore carry a larger share of fitted score variation. The relevant question is not whether all signal remains structured, but whether `findr` can still identify the part that is structured and interpretable. The coefficient paths

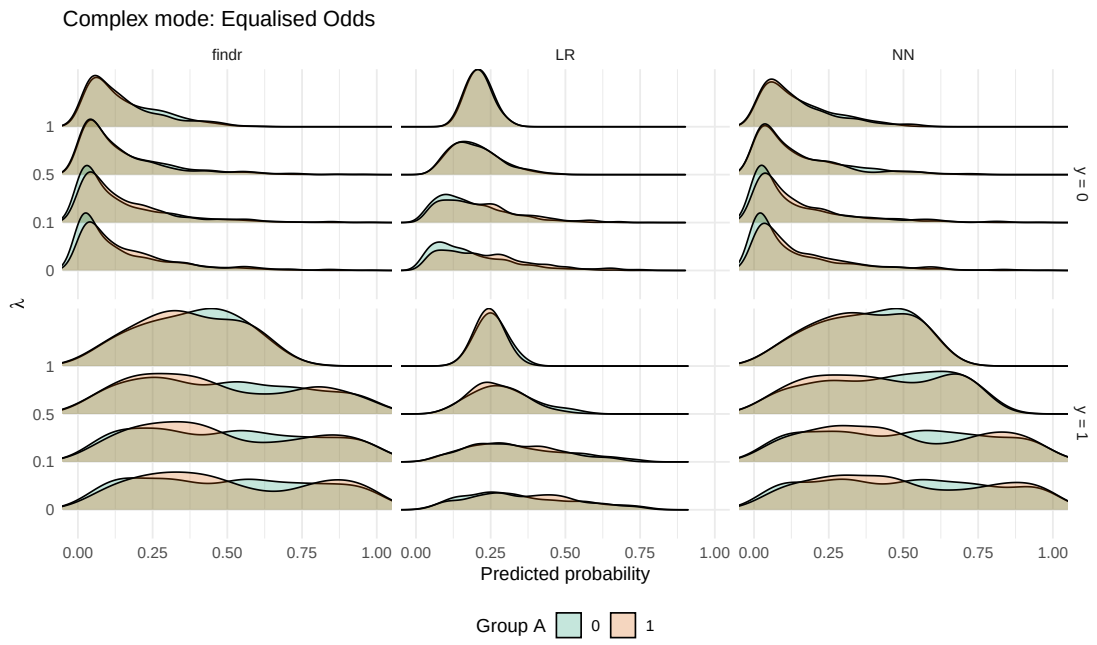

Figure 5: Score distributions in the complex simulation setting. The panels show fitted score distributions across groups and outcome strata as λ varies.

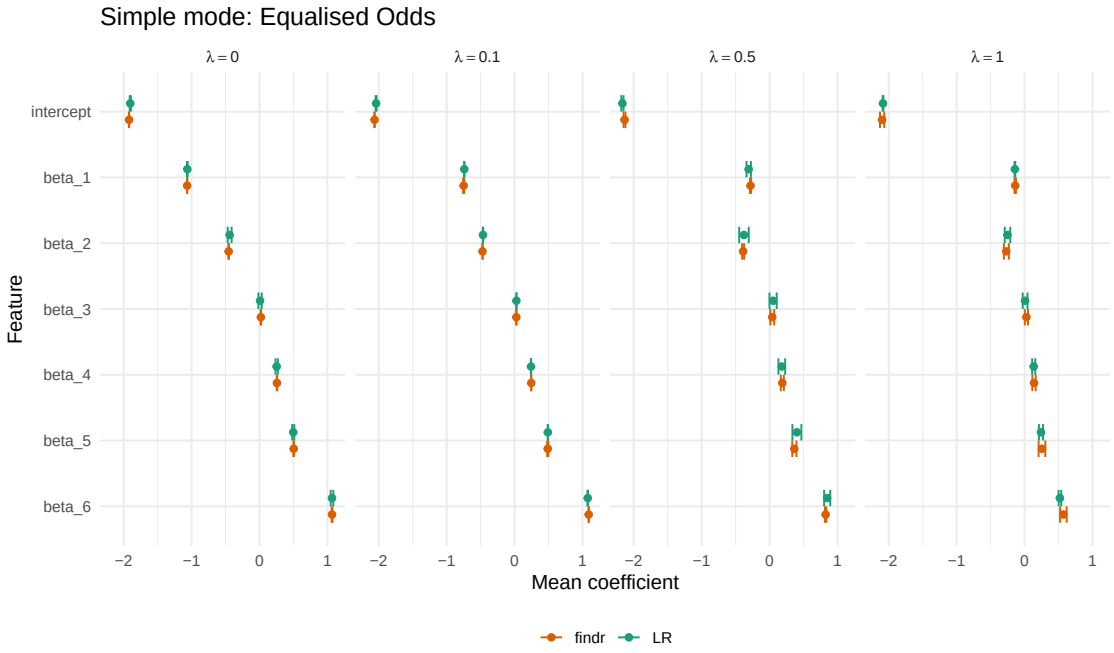

Figure 6: Estimated structured coefficients in the simple simulation setting. The true signal is linear, and the structured coefficients from `findr` follow the same coefficient pattern as LR.

in Figure 7 remain coherent with LR, which indicates that the structured component still captures the linear part of the signal. At the same time, Table 2 shows that EVR is lower than in the simple mode, ranging from about 0.44 to 0.56 across the reported λ values. This means that a substantial share of logit-scale variation is assigned to the neural residual rather than to the structured component. DDR also becomes nonzero, so some decisions depend on the

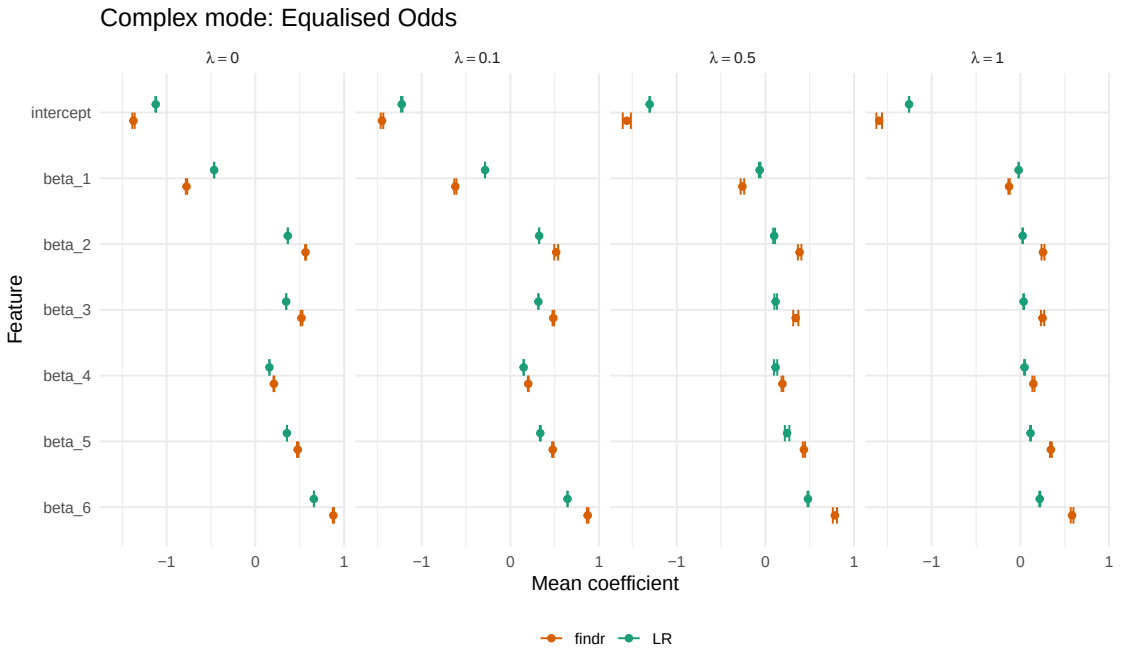

Figure 7: Estimated structured coefficients in the complex simulation setting. The structured coefficients from `findr` remain coherent with `LR`, while the neural residual accounts for nonlinear terms and interactions.

integrated model rather than the structured component alone. CSER remains small for some features but can increase for the most affected feature as λ grows, indicating that fairness regularisation and nonlinear residual structure can alter local directional behaviour. This is also shown in Figure 7, where the effect of increases in λ is particularly pronounced for β_1. Together, these results show that `findr` retains coefficient interpretations coherent with the logistic baseline while also quantifying how much fitted variation is carried by the structured component and how much is carried by the residual. A standard logistic model cannot make this split because it has no residual component, and a pure neural network cannot make it because it has no separate interpretable component.

Table 2: **Complex mode:** Equalised Odds. EVR, DDR, CSER (min) and CSER (max) reported as mean (SD) for the findr model by λ.

λ	EVR	DDR	CSER (min)	CSER (max)
0.0	0.556 (0.013)	0.123 (0.004)	0.000 (0.000)	0.038 (0.007)
0.1	0.525 (0.006)	0.120 (0.001)	0.000 (0.000)	0.055 (0.013)
0.5	0.436 (0.016)	0.112 (0.004)	0.000 (0.000)	0.289 (0.032)
1.0	0.469 (0.017)	0.068 (0.003)	0.011 (0.006)	0.424 (0.020)

5 Application

To complement the results of the simulation study, we proceed with comparing `findr` and its two benchmarks across real-world lending data.

5.1 Datasets and sensitive attribute

We study eight public datasets frequently used in credit-scoring modelling. All datasets use *age* as the sensitive attribute (denoted as A in Section 3), binarised at 25 ("younger" vs. "older"), which has been a common setting in previous work (Kozodoi et al., 2022; Kim et al., 2023; Kamiran and Calders, 2009). Below, we report the sample size n, number of features, default rate, and sensitive-class share (proportion with age ≤ 25) for each dataset, with further details in Appendix B.

South German Credit. $n = 1{,}000$ applicants with 19 features; default rate 30.0%; sensitive class share 19.0%.

Taiwan Credit Default. $n = 30{,}000$ clients with 22 features; default rate 22.12%; sensitive class share 12.9%.

Give Me Some Credit. $n = 150{,}000$ applicants with 9 features; default rate 6.68%; sensitive class share 2.02%.

Vehicle Car Loan. $n = 233{,}154$ loans with 32 features; default rate 21.71%; sensitive class share 22.84%.

MyHome Loan. $n = 7{,}000$ loans with 7 features; default rate 40.0%; sensitive class share 7.03%.

Thomas Loan. $n = 1{,}225$ applicants with 13 features; default rate 26.37%; sensitive class share 7.67%.

UK Credit. $n = 30{,}000$ applicants with 13 features; default rate 4.0%; sensitive class share 19.75%.

PAKDD 2010. $n = 50{,}000$ applicants with 51 features; default rate 26.08%; sensitive class share 11.49%.

We split the data into training, validation, and test sets using a 70/10/20 split. Continuous features are standardised, and categorical variables are encoded using weights of evidence. We train `LR`, `NN`, and `findr` across a λ grid ranging from 0 to 2 and repeat each configuration over 10 random seeds. Hyperparameters are selected on the validation split, with details given in Appendix C, and all reported metrics are evaluated out of sample on the test split. As in the simulation study, the main text reports equalised-odds results. The other fairness criteria were also examined, but they did not add new qualitative insights to the model comparisons.

5.2 Results

5.2.1 Out-of-sample performance

Figure 8 summarises out-of-sample discrimination and calibration across the λ path, measured by AUC and Brier score. The dashed horizontal line gives the unpenalised CatBoost result (Prokhorenkova et al., 2018), included as a state-of-the-art tabular-data reference that is neither fairness-constrained nor interpretable. It provides a scale for assessing how close the three penalised model classes remain to a strong predictive benchmark as λ varies.

The plotted paths show that the predictive effect of the fairness penalty is not uniform across datasets. In some datasets, AUC and Brier score remain relatively stable along the path, while in others stronger penalisation moves the fitted scores toward weaker discrimination or poorer calibration.

Tables 3 and 4 give two cuts of these paths, at $\lambda = 0$ and $\lambda = 0.1$, to make the contrast between empirical predictive regimes easier to read. Taiwan and GMSC are the most distinct cases, which we treat as nonlinear regimes because the flexible models improve substantially over logistic regression. At $\lambda = 0$, the neural model gives AUC gains of about 5.5 and 18.7 percentage points, respectively. `findr` recovers most of this gain, reaching AUC 0.774 in Taiwan and 0.850 in GMSC. The Brier scores follow the same pattern, with the flexible models giving lower prediction error than the linear logistic case. In the other six datasets, the three models are much closer at $\lambda = 0$, suggesting approximately linear regimes where the linear baseline already captures most of the predictive signal. This interpretation is further supported by the structured-component diagnostics in Subsection 5.2.3.

The full λ paths (Figure 8) add context beyond the unpenalised cut. Taiwan and Vehicle Loan show the largest predictive cost at high penalties. At $\lambda = 1$, AUC falls from 0.774 to 0.665 for `findr` in Taiwan and from 0.641 to 0.513 in Vehicle Loan. The neural and logistic paths show similar deterioration in Vehicle Loan, where the high-penalty solutions approach nearly uninformative discrimination. GMSC behaves differently. The flexible models lose AUC as λ increases, but they remain well above the logistic baseline through moderate penalties. This indicates that the nonlinear signal remains useful even after adding the fairness penalty.

The remaining datasets show milder or more metric-specific changes. South German and UK Credit retain relatively stable AUC across the path, although South German shows a visible increase in Brier score at larger penalties. Thomas shows modest Brier changes but lower AUC under stronger penalisation. PAKDD 2010 is the main exception to a simple performance-loss description. Its AUC declines as λ increases, but its Brier score improves for all three models, suggesting that the penalty reduces discrimination while producing less extreme probability forecasts. In summary, Figure 8 shows that `findr` follows the neural model when nonlinear predictive structure is useful and stays close to logistic regression in approximately linear regimes.

Although not the focus of this paper, we note that Figure 8 also supports the focus on (semi-)parametric model structures. The unconstrained `findr`, `LR`, and `NN` models typically perform close to CatBoost, indicating that alternative model classes (e.g., tree-based ensemble models) would not provide substantially higher predictive performance.

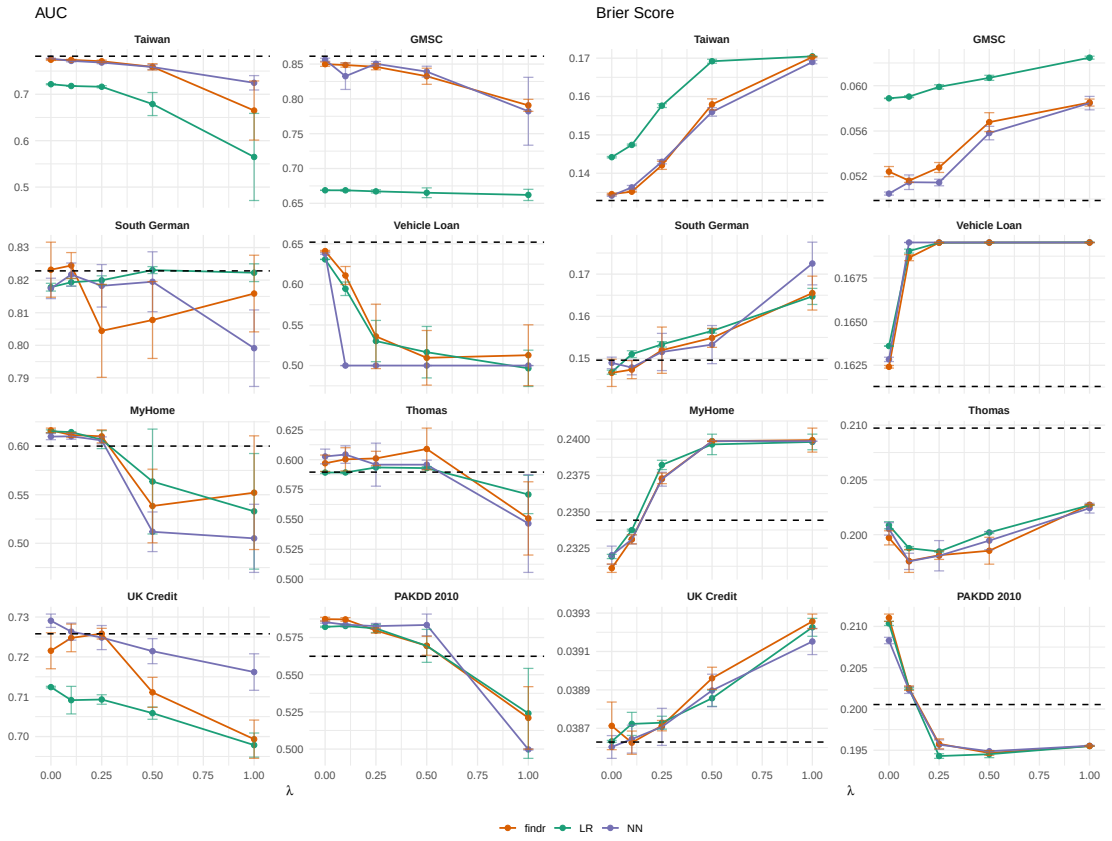

Figure 8: Out-of-sample discrimination and calibration across datasets under the equalised-odds penalty. Curves show AUC and Brier score as the fairness multiplier λ varies. Dashed horizontal lines show the unpenalised CatBoost reference.

Table 3: Out-of-sample AUC and Brier score for the unpenalised models. Bold marks the best result per metric within each row. Datasets are grouped by predictive regime.

	$\lambda = 0$					
	AUC (↑)			Brier (↓)		
Dataset	LR	NN	`findr`	LR	NN	`findr`
Nonlinear regime						
Taiwan	0.722	**0.777** (0.001)	0.774 (0.002)	0.144	**0.134** (0.000)	0.135 (0.000)
GMSC	0.669	**0.856** (0.002)	0.850 (0.003)	0.059	**0.050** (0.000)	0.052 (0.000)
Linear regime						
South German	0.818	0.817 (0.003)	**0.823** (0.008)	0.147	0.149 (0.001)	**0.147** (0.003)
Vehicle Loan	0.631	0.638 (0.001)	**0.641** (0.001)	0.164	0.163 (0.000)	**0.162** (0.000)
MyHome	0.615	0.610 (0.003)	**0.616** (0.002)	0.232	0.232 (0.001)	**0.231** (0.000)
Thomas	0.589	**0.603** (0.006)	0.597 (0.007)	0.201	0.201 (0.001)	**0.200** (0.001)
UK Credit	0.712	**0.729** (0.002)	0.722 (0.005)	0.039	**0.039** (0.000)	0.039 (0.000)
PAKDD 2010	0.582	0.585 (0.001)	**0.587** (0.001)	0.210	**0.208** (0.000)	0.211 (0.000)

5.2.2 Fairness

Each panel in Figure 9 corresponds to a dataset, showing the score-level predictor frontiers and model paths as λ varies.

Table 4: Out-of-sample AUC and Brier score for the three models at $\lambda = 0.1$. Bold marks the best result per metric within each row. Datasets are grouped by predictive regime.

	$\lambda = 0.1$					
	AUC (↑)			Brier (↓)		
Dataset	LR	NN	`findr`	LR	NN	`findr`
Nonlinear regime						
Taiwan	0.718	0.772 (0.002)	**0.774** (0.001)	0.147	0.136 (0.001)	**0.135** (0.000)
GMSC	0.669	0.833 (0.019)	**0.849** (0.003)	0.059	**0.051** (0.001)	0.052 (0.000)
Linear regime						
South German	0.819	0.822 (0.004)	**0.824** (0.004)	0.151	0.148 (0.002)	**0.147** (0.002)
Vehicle Loan	0.595	0.500 (0.000)	**0.611** (0.011)	0.169	0.170 (0.000)	**0.169** (0.000)
MyHome	**0.614**	0.610 (0.003)	0.611 (0.003)	0.234	**0.233** (0.000)	0.233 (0.000)
Thomas	0.589	**0.604** (0.007)	0.600 (0.010)	0.199	**0.198** (0.001)	0.198 (0.001)
UK Credit	0.709	**0.726** (0.002)	0.725 (0.003)	0.039	0.039 (0.000)	**0.039** (0.000)
PAKDD 2010	0.583	0.584 (0.001)	**0.587** (0.001)	0.202	**0.202** (0.000)	0.203 (0.000)

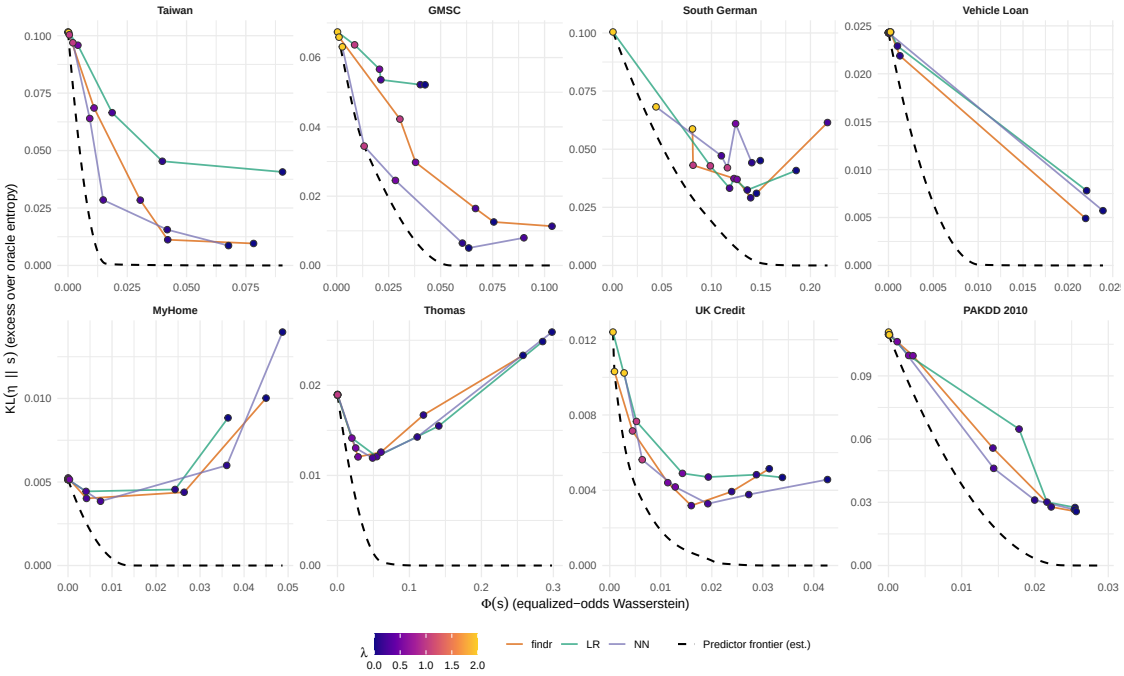

Figure 9: Score-level predictor frontiers and fitted model paths for the application datasets under equalised odds. The horizontal axis is the empirical equalised-odds Wasserstein gap, and the vertical axis is excess expected log loss relative to the reference risk. The colour of each point indicates the value of λ.

The frontier comparison shows how each model moves through the score-level accuracy–fairness trade-off as the penalty increases. The main pattern is heterogeneous across datasets. In the nonlinear regimes, `findr` tends to remain close to the flexible model and closer to the frontier than logistic regression. In the approximately linear regimes, its path is usually comparable to the simpler alternatives and is not systematically worse. The main exception is South German, where the logistic path reaches the low-gap frontier region more directly than `findr` and `NN`.

The individual panels support this pattern in different ways. Thomas and South German have the widest fairness scales, but they do not show the same trade-off. In Thomas, all three models reduce the fairness gap with little visible increase in excess loss. South German is more model-dependent, with logistic regression moving more directly toward the low-gap region. These two datasets are also the smallest datasets (see Appendix B), which may contribute to the wider and less stable fitted paths because fairness distance is estimated within outcome-by-group cells.

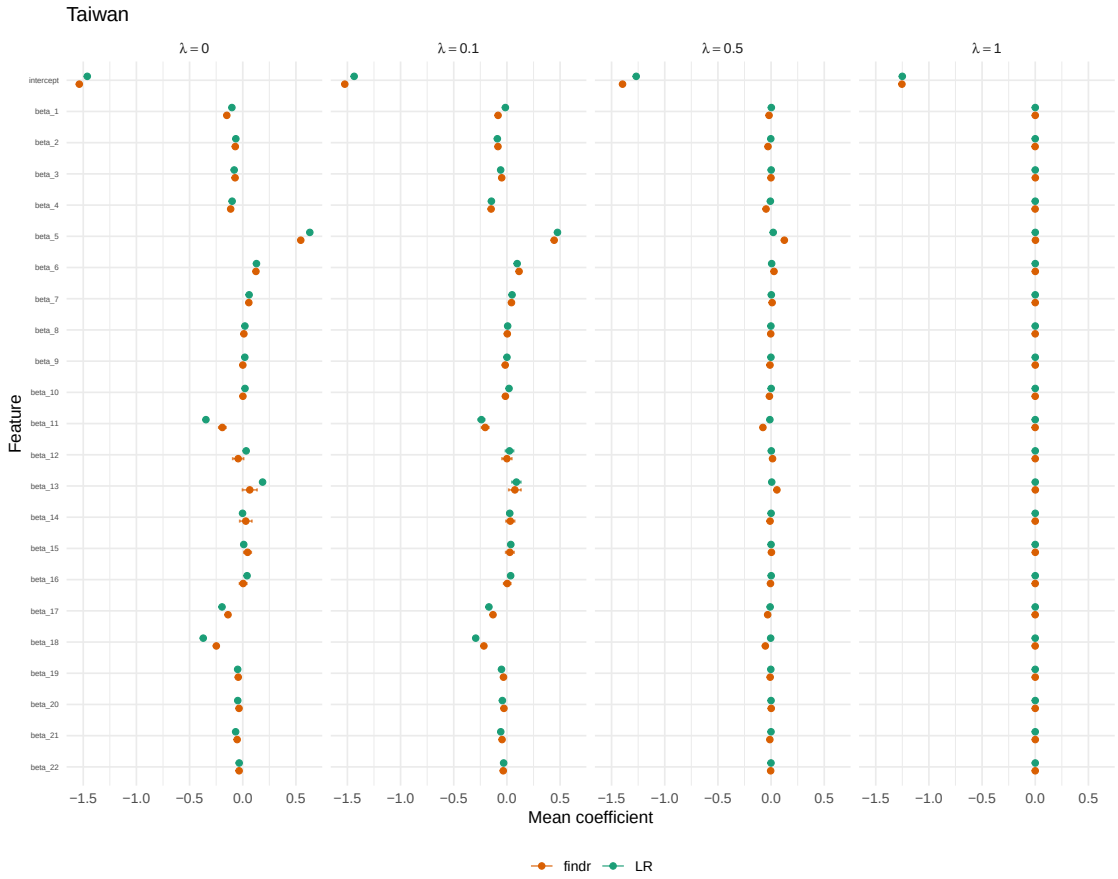

Figure 10: Taiwan Credit Default structured coefficients across seeds under equalised odds for selected values of λ.

Taiwan and GMSC are the nonlinear predictive regimes identified in Subsection 5.2.1. In both panels, the flexible paths are closer to the estimated frontier than logistic regression over the whole path. This is most visible in GMSC, where `NN` and `findr` begin with much lower excess loss than logistic regression at moderate fairness gaps. In Taiwan, all models can drive the fairness gap close to zero, but with a large increase in excess loss. For `findr`, the gap falls from approximately 0.08 at $\lambda = 0$ to nearly zero at $\lambda = 2$, while excess loss rises from about 0.01 to 0.10. GMSC has a similar leftward movement, but the flexible paths retain a clearer advantage at lower and moderate penalties. This is consistent with the performance results showing that nonlinear signal is useful in these two datasets.

Vehicle Loan, UK Credit, MyHome, and PAKDD 2010 have smaller fairness scales than Thomas, South German, Taiwan, and GMSC. Vehicle Loan and UK Credit show near-zero gaps at high λ with moderate vertical movement. MyHome is also compact, with gaps below about 0.05, and all three model paths move close to the low-gap region. PAKDD 2010 shows a different pattern. Its initial gaps are already small, near 0.025, but forcing them close to zero produces one of the largest excess-loss increases in the figure. This cautions against reading all predictive summaries in the same way. The frontier's log-loss scale shows a cost, whereas Tables 3 and 4 show improved Brier scores at the reported λ values.

5.2.3 Interpretability

The structured part of `findr` gives a coefficient vector that can be read in the same way as in logistic regression. Figures 10 and 11 show this information for Taiwan and MyHome, which represent a nonlinear and an approximately linear regime, respectively. Analogous coefficient plots are obtained for the other datasets. Since the coefficients are displayed in generic feature notation, we focus on their sign, magnitude, and stability rather than on feature-specific substantive interpretation. In both examples, the structured coefficients remain readable across the selected values of λ, even though the full model also includes a neural residual.

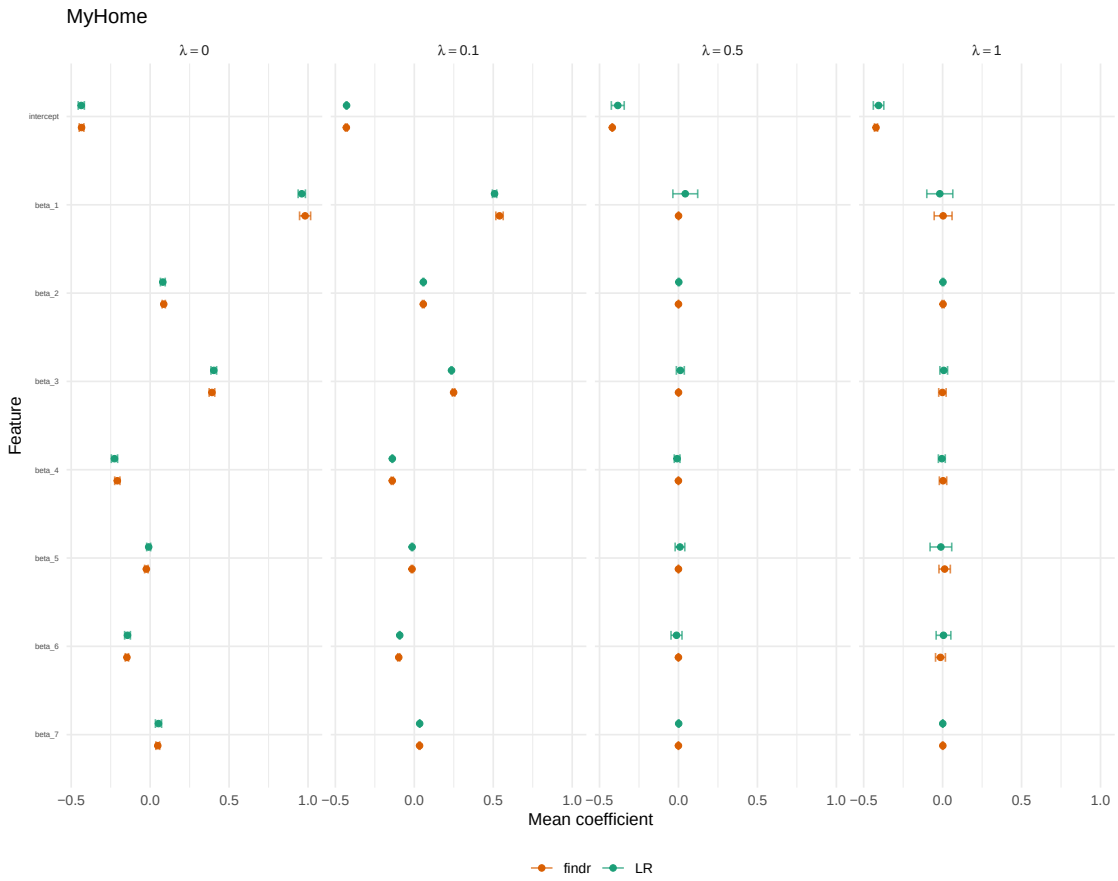

Figure 11: MyHome Loan structured coefficients across seeds under equalised odds for selected values of λ.

Table 5: EVR, DDR, and CSER for `findr` at $\lambda = 0$ under equalised odds. Entries are mean and standard deviation across seeds. Datasets are grouped by predictive regime.

	$\lambda = 0$			
Dataset	EVR	DDR	Min CSER	Max CSER
Nonlinear regime				
Taiwan	0.527 (0.008)	0.084 (0.002)	0.088 (0.050)	0.470 (0.046)
GMSC	0.399 (0.034)	0.034 (0.004)	0.018 (0.010)	0.215 (0.195)
Linear regime				
South German	0.957 (0.017)	0.030 (0.013)	0.000 (0.000)	0.158 (0.126)
Vehicle Loan	0.805 (0.014)	0.004 (0.001)	0.001 (0.001)	0.204 (0.017)
MyHome	0.935 (0.024)	0.022 (0.006)	0.000 (0.000)	0.004 (0.004)
Thomas	0.981 (0.015)	0.005 (0.006)	0.000 (0.000)	0.051 (0.071)
UK Credit	0.749 (0.024)	0.000 (0.000)	0.051 (0.016)	0.429 (0.012)
PAKDD 2010	0.946 (0.008)	0.028 (0.003)	0.000 (0.000)	0.044 (0.032)

We are also interested in whether these coefficients are enough to explain the fitted decisions. As in the simulation study, a readable coefficient vector does not imply that the linear component describes the whole fitted risk surface. Tables 5 and 6 report this comparison for $\lambda = 0$ and $\lambda = 0.1$. Recall that EVR measures how much logit-scale variation is assigned to the structured component, DDR measures decision agreement between the structured component and the full model, and CSER measures whether local coefficient directions remain consistent after adding the residual.

Table 6: EVR, DDR, and CSER for findr at $\lambda = 0.1$ under equalised odds. Entries are mean and standard deviation across seeds. Datasets are grouped by predictive regime.

	$\lambda = 0.1$			
Dataset	EVR	DDR	Min CSER	Max CSER
Nonlinear regime				
Taiwan	0.505 (0.009)	0.095 (0.002)	0.039 (0.028)	0.462 (0.034)
GMSC	0.392 (0.019)	0.021 (0.002)	0.011 (0.012)	0.145 (0.040)
Linear regime				
South German	0.965 (0.011)	0.033 (0.013)	0.000 (0.000)	0.147 (0.085)
Vehicle Loan	0.916 (0.020)	0.000 (0.000)	0.000 (0.000)	0.193 (0.124)
MyHome	0.876 (0.036)	0.008 (0.004)	0.007 (0.021)	0.067 (0.054)
Thomas	0.952 (0.033)	0.002 (0.003)	0.000 (0.000)	0.093 (0.101)
UK Credit	0.753 (0.014)	0.000 (0.000)	0.016 (0.011)	0.401 (0.008)
PAKDD 2010	0.927 (0.005)	0.022 (0.002)	0.000 (0.000)	0.120 (0.058)

In the approximately linear regime datasets, the coefficients are close to sufficient for describing the fitted decisions. At $\lambda = 0$, EVR is 0.957 in South German, 0.935 in MyHome, 0.981 in Thomas, and 0.946 in PAKDD 2010. DDR is also small in these datasets, so the structured component usually gives the same binary decision as the full model. These are the cases where the interpretation is closest to logistic regression, with the additional check that little decision-relevant information is left to the residual. Vehicle Loan and UK Credit are more intermediate. Their EVR values are lower, 0.805 and 0.749, but DDR remains close to zero. The residual therefore contributes to continuous score variation, with limited effect on thresholded decisions.

Taiwan and GMSC show why having coefficients is not the same as having a complete coefficient-based explanation. Their EVR values are much lower, 0.527 and 0.399 at $\lambda = 0$ and 0.505 and 0.392 at $\lambda = 0.1$, which means that the neural residual carries a substantial share of logit-scale variation. In Taiwan, nonzero DDR and high maximum CSER suggest that the residual can affect both some decisions and some local directional interpretations. In GMSC, DDR remains low despite low EVR, so the residual changes the continuous score surface more than the final thresholded decisions.

Read together with the coefficient plots, these results clarify the role of the structured component. In MyHome, the plotted coefficients are close to a full decision-level explanation. In Taiwan, they remain meaningful, but the residual must be examined to understand the full fitted model. In this way, findr keeps the familiar coefficient interpretation when it is adequate, while the proposed diagnostics indicate when additional nonlinear information must also be considered.

6 Conclusion

In high-stakes decisions such as credit scoring, there is a common tension between performance, interpretability, and fairness. Linear models provide coefficients that are easy to interpret and audit, but they can miss nonlinear structure or higher-order interactions. Flexible models can accommodate these predictive effects, but they usually do not by themselves identify which part of the fitted score is interpretable, nor whether structured explanations are sufficient for decisions.

We introduce findr, a semi-structured approach for binary regression that combines a readable linear logit, an orthogonal neural residual, and an in-processing Wasserstein fairness penalty. The model is interpretable through its coefficient component, flexible through its neural residual, and fairness-aware through a penalty that compares score distributions during training. The additive decomposition is useful beyond the specific linear specification used in the empirical analysis. It can identify a structured and interpretable component, which could be replaced by more expressive structured models such as GAMs, while keeping the residual component available for information that is difficult to represent in a standard structured model. This also opens the possibility of incorporating richer data sources, such as text, through the neural residual. The flexibility of the fairness penalisation allows us to use the same construction with different fairness notions by changing which conditional score distributions enter the Wasserstein distance.

We also contribute tools for evaluating and comparing the fitted models. The interpretability diagnostics separate three questions. EVR measures how much logit-scale variation is assigned to the structured component. DDR measures whether the structured and full logits lead to the same binary decisions. CSER identifies cases where the direction

implied by a specific coefficient differs from the full model's local response. The score-level frontiers provide an empirical benchmark for how much predictive loss is associated with lower fairness gaps on the same validation sample.

We illustrate the approach in a simulation study. When the true signal is linear, findr behaves like logistic regression, with most fitted variation assigned to the structured component and little disagreement between the structured and full decisions. When the signal contains nonlinearities or interactions, the residual carries a larger share of the logit-scale variation, while the structured coefficients remain coherent with the linear part of the signal. The Wasserstein penalty moves scores toward lower fairness gaps by acting on entire score distributions rather than on a single threshold or a single distributional summary.

We apply the approach to eight credit-related datasets. The results are more heterogeneous, as expected in real applications. In Taiwan and GMSC, the nonlinear predictive component is relevant, and findr recovers much of the gain of the neural model while retaining a structured coefficient component. In other datasets, the linear baseline already captures most of the predictive signal. We also show that fairness costs vary across datasets and across evaluation criteria. Some datasets allow large reductions in the Wasserstein gap with little excess log-loss cost, while others require stronger changes to the fitted scores. We then use the interpretability diagnostics to clarify when the coefficient component is close to a full decision-level explanation and when the residual must also be examined.

Overall, our results support semi-structured modelling as a practical middle ground between logistic regression and fully flexible neural models. We do not remove the need to choose a fairness criterion or to evaluate model behaviour in context. We do, however, make the performance, fairness, and interpretability trade-offs more explicit. In settings where the linear explanation is adequate, findr behaves close to the transparent baseline. In settings where nonlinear structure matters, it can use that structure while reporting how much decision-relevant information remains outside the coefficient component.

Several extensions follow from this framework. On the fairness side, future work could extend the Wasserstein penalties to predefined multi-group settings, intersectional group definitions, and individual-based notions such as counterfactual fairness. On the interpretability side, the diagnostics suggest a route toward variable selection under an explicit notion of interpretability, where the goal is to retain the variables that carry most of the interpretable signal while leaving remaining predictive information to the residual.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, 34:4699–4711.

Baesens, B., Goethals, A., Lessmann, S., De Vos, S., Bravo, C., Martens, D., Medina-Olivares, V., Mues, C., Oskarsdóttir, M., Broucke, S. v., Verdonck, T., and Verbeke, W. (2026). Foundation models for credit risk prediction: A game changer? *arXiv preprint arXiv:2605.18147*.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Chang, C.-H., Caruana, R., and Goldenberg, A. (2022). NODE-GAM: Neural generalized additive model for interpretable deep learning.

Chen, P., Wu, L., and Wang, L. (2023). Ai fairness in data management and analytics: A review on challenges, methodologies and applications. *Applied sciences*, 13(18):10258.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Chuang, C.-Y. and Mroueh, Y. (2021). Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*.

Consumer Financial Protection Bureau (2026). 12 CFR Part 1002: Equal Credit Opportunity Act (Regulation B).

Dakovic, R., Czado, C., and Berg, D. (2010). Bankruptcy prediction in norway: a comparison study. *Applied economics letters*, 17(17):1739–1746.

Dastile, X., Celik, T., and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263.

De Vos, S., Van Belle, J., Algaba, A., Verbeke, W., and Verboven, S. (2025). Decision-centric fairness: Evaluation and optimization for resource allocation problems. *arXiv preprint arXiv:2504.20642*.

Djeundje, V. B. and Crook, J. (2019). Identifying hidden patterns in credit risk survival data using generalised additive models. *European Journal of Operational Research*, 277(1):366–376.

Dumitrescu, E., Hué, S., Hurlin, C., and Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3):1178–1192.

European Banking Authority (2020). Guidelines on loan origination and monitoring.

European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

Frosst, N. and Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.

Gramegna, A. and Giudici, P. (2021). Shap and lime: an evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, 4:752558.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, volume 43. CRC Press.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., and Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52.

Hurlin, C., Pérignon, C., and Saurin, S. (2026). The fairness of credit scoring models. *Management Science*.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2020). Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR.

Kamiran, F. and Calders, T. (2009). Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6, Karachi, Pakistan. IEEE.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 35–50. Springer.

Kim, S., Lessmann, S., Andreeva, G., and Rovatsos, M. (2023). Fair models in credit: Intersectional discrimination and the amplification of inequity. *arXiv preprint arXiv:2308.02680*.

Korangi, K., Mues, C., and Bravo, C. (2023). A transformer-based model for default prediction in mid-cap corporate markets. 308(1):306–320.

Kozodoi, N., Jacob, J., and Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., and Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 623–631. Association for Computing Machinery.

Luber, M., Thielmann, A. F., and Säfken, B. (2023). Structural neural additive models: Enhanced interpretable machine learning. *arXiv preprint arXiv:2302.09275*.

Martens, D., Baesens, B., Van Gestel, T., and Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, 183(3):1466–1476.

Medina-Olivares, V., Lessmann, S., and Klein, N. (2024). The deep promotion time cure model. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):18848–18858.

Medina-Olivares, V., Xia, W., Lessmann, S., and Klein, N. (2026). Semi-structured multi-state delinquency model for mortgage default. *arXiv preprint arXiv:2603.26309*.

Menon, A. K. and Williamson, R. C. (2018). The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431.

Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31.

Rügamer, D., Kolb, C., and Klein, N. (2024). Semi-structured distributional regression. *The American Statistician*, 78(1):88–99.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR.

Stevenson, M., Mues, C., and Bravo, C. (2021). The value of text for small business default prediction: A Deep Learning approach. *European Journal of Operational Research*, 295(2):758–771.

Tan, S., Hooker, G., Koch, P., Gordo, A., and Caruana, R. (2023). Considerations when learning additive explanations for black-box models. *Machine Learning*, 112(9):3333–3359.

Taskesen, B., Nguyen, V. A., Kuhn, D., and Blanchet, J. (2020). A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*.

Thielmann, A. F., Kruse, R.-M., Kneib, T., and Säfken, B. (2024). Neural additive models for location scale and shape: A framework for interpretable neural regression beyond the mean. In *International Conference on Artificial Intelligence and Statistics*, pages 1783–1791. PMLR.

Villani, C. et al. (2008). *Optimal transport: old and new*, volume 338. Springer.

Wan, M., Zha, D., Liu, N., and Zou, N. (2023). In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., and Jordan, M. (2020). Robust optimization for fairness with noisy protected groups. *Advances in Neural Information Processing Systems*, 33:5190–5203.

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

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

A Computational construction of the predictor frontier

This appendix summarises the numerical construction used for the score-level frontier in Section 3.5. For a score vector s, the code evaluates $\widehat{\Phi}(s; y, a)$ by exact step-function integration of the empirical CDF gaps within the realised validation cells $(Y = y, A = a)$. It evaluates $\widehat{R}_{\log}(r, s)$ by passing r as the first argument to the log-loss routine, so the loss is the expected Bernoulli log loss under the reference risk rather than the realised validation log loss.

For each frontier multiplier ξ, the scalarised objective is

$$Q_\xi(s) = \widehat{R}_{\log}(r, s) + \xi \widehat{\Phi}(s; y, a).$$

The default grid is $\{0\} \cup \text{logspace}(-2, 2, 40)$. The case $\xi = 0$ is returned in closed form as $s = r$. For $\xi > 0$, the solver starts from the candidate in the pool consisting of r, optional fitted-model score vectors, and the previous multiplier solution that gives the smallest value of Q_ξ.

The iteration uses a soft-cell approximation to the fairness direction. For cells (u, g), define

$$\pi_{ug}(x_i) = \mathbf{1}(a_i = g)\{r_i \mathbf{1}(u = 1) + (1 - r_i)\mathbf{1}(u = 0)\}, \qquad p_{ug} = n^{-1} \sum_i \pi_{ug}(x_i),$$

and let $F_{ug}(v; s)$ be the corresponding weighted empirical CDF of the current scores. The effective sign used in the update is

$$\Theta_i = \sum_{u\in\{0,1\}} \sum_{g\in\{0,1\}} \frac{\pi_{ug}(x_i)}{p_{ug}} \operatorname{sign}\{F_{ug}(s_i; s) - F_{u,1-g}(s_i; s)\}. \tag{15}$$

The score update is

$$s_i^{(t+1)} = (1-\rho_\xi)s_i^{(t)} + \rho_\xi \left\{ r_i + \xi s_i^{(t)}(1-s_i^{(t)})\Theta_i^{(t)} \right\}, \qquad \rho_\xi = \frac{\rho}{1+\xi}, \tag{16}$$

followed by clipping to $[\varepsilon, 1-\varepsilon]$. A positive Θ_i increases the next score for observation i. Since larger scores mean higher predicted default risk, this direction raises the default probability assigned to observations whose soft outcome-cell memberships place them in cells with relatively larger weighted CDF mass at their current score. It moves mass to the right in those cells. The objective used to accept iterates is still the realised-cell objective Q_ξ, so this update should be read as a numerical heuristic rather than an exact subgradient step for $\widehat{\Phi}(s; y, a)$.

The solver retains the best objective value among the initial score, periodic iterates, and the Polyak–Ruppert average of the second half of the path. Candidate scores are then added to the point set before envelope extraction, because they are feasible score vectors on the same validation sample. If enabled, one refinement pass inserts up to fifteen additional multipliers where adjacent solver-produced points have large normalised chord lengths. The current implementation chooses these new multipliers by log-midpoints, with the special case $\xi_0 = 0 < \xi_1$ using $\xi_1/2$. The final frontier sorts all points by $\widehat{\Phi}$ and takes the running minimum of $\widehat{R}_{\log}$, collapsing duplicate unfairness values.

B Datasets summary statistics

Table 7: Dataset summary statistics. Default rates are reported overall and by sensitive group ($A = 0$ vs. $A = 1$).

Dataset	n	Default rate (%)	Sensitive share (%)	Default rate A=0 (%)	Default rate A=1 (%)
South German	1,000	30.00	19.00	27.16	42.11
Taiwan	30,000	22.12	12.90	21.45	26.66
GMSC	150,000	6.68	2.02	6.59	11.16
Vehicle loan	233,154	21.71	22.84	21.03	24.00
MyHome	7,000	40.00	7.03	39.67	44.31
Thomas	1,225	26.37	7.67	24.49	48.94
UK	30,000	4.00	19.75	3.55	5.84
PAKDD	50,000	26.08	11.49	25.02	34.29

C Hyperparameter search

Table 8 reports the hyperparameter space used for the empirical experiments. We use discrete random search separately for each value of λ. Candidate configurations are ranked by their mean validation objective over two optimisation restarts. Training is limited to 6,000 epochs, with early stopping after 20 epochs without an improvement during tuning. All models are trained with AdamW and binary cross-entropy with logits. After selecting a configuration, each experiment is repeated over 10 random seeds and evaluated on the held-out test set.

Table 8: Hyperparameter search space.

Hyperparameter	Candidate values
Learning rate	$\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$
Weight decay	$\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$
Batch size	$\{32, 64, 128, 256\}$
Hidden-layer widths	$\{[4], [16, 8], [64, 32], [128, 64]\}$
Activation	{ReLU, GELU, tanh}
Dropout rate	$\{0, 0.1, 0.25\}$