Machine learning optimization approaches for fair lending in finance

- Alexey Miroshnikov
- Data Science Research Group, Discover Financial Services

University of North Carolina – Charlotte
Center for Trustworthy Artificial Intelligence in Model Risk Management, February 26, 2025
(slides amended April, 2025)

1. Introduction

Fair Lending

ML models and strategies that rely on ML models are subject to federal laws and regulations, including the Equal Credit Opportunity Act (ECOA), Fair Housing Act (FHA), and Equal Employment Opportunity Act (EEOA).

ECOA and FHA laws and regulations prohibit discrimination against protected classes (sub-populations) in lending; thus, disparities against the sub-populations must be considered.

EEOA laws forbid discrimination against employees and job applicants on the bases of race, color, religion, sex, national origin, disability, or age.

Examples of protected attributes: race, gender, age, ethnicity, national origin, marital status, etc.

Explainability

ML risk models use historical, consumer and consumer reporting information to estimate the probability of default. US Federal regulations require lenders to provide applicants with the primary factors that contribute to an adverse action (i.e., decline).

Explainability methods. There are a variety of mathematical and statistical techniques that quantify the contribution of each element from the input vector to the predictive model output given the distribution of inputs. Game theoretic approaches are popular, as well as models that are inherently interpretable.

Notation

Distribution (X, G, Y)

- $X = (X_1, \dots, X_n) \in \mathbb{R}^n$, predictors
- $G \in \{0,1\}$ (e.g. male/female)
- $Y \in \{0,1 = \text{adverse action}\}$, response variable

Models

- $f(X;\theta) = \widehat{\mathbb{P}}(Y = 1|X)$, a trained classification score, $\theta \in \mathbb{R}^m$
- $\widehat{Y_t} = 1_{\{f(X)>t\}}$, a classifier for a given threshold $t \in [0,1]$

Note: sometimes, we work with a raw probability score $f(X;\theta) := logit(\widehat{\mathbb{P}}(Y = 1|X))$.

Local fairness metrics for classifiers

ML bias can be viewed as an ability to differentiate between subpopulations at the level of data or outcomes *[Dwork et al 2012]*

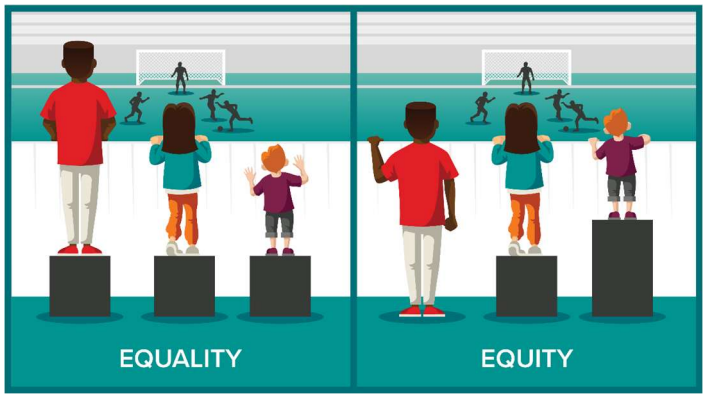
- Statistical parity, $\hat{Y}_t \in \{0,1\}$ [Feldman et al, 2015]

$$\text{Bias}^{st}(\hat{Y}_t|\text{G}) = \left|\mathbb{P}(\hat{Y}_t = 0|G = 0) - \mathbb{P}(\hat{Y}_t = 0|G = 1)\right|$$

- Equalized odds, $\hat{Y}_t \in \{0,1\}$ [Hardt et al, 2015]

$$\text{Bias}^{eo}(\hat{Y}_t|\text{G}) = \left|\mathbb{P}(\hat{Y} = 0|Y = y, G = 0) - \mathbb{P}(\hat{Y} = 0|Y = y, G = 1), y \in \{0,1\}\right|$$

For generalizations, see [Miroshnikov-Kotsiopoulos-Ravi Kannan, ML Springer (2022)]



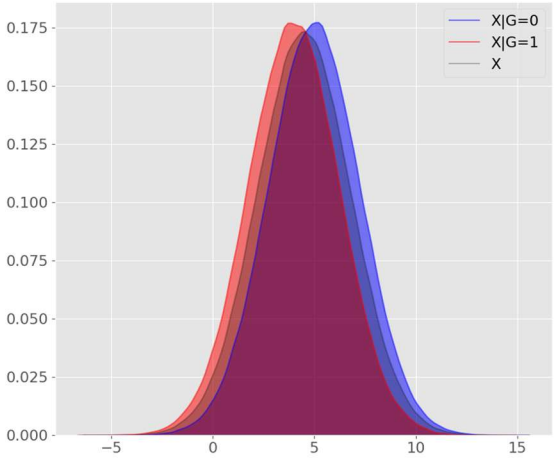
Performance-Fairness trade-off example

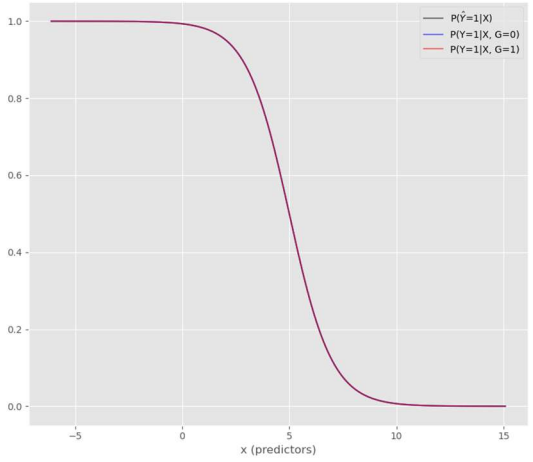
Statistical parity classifier bias

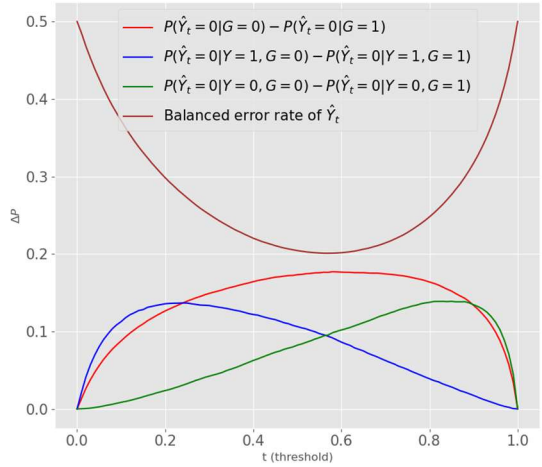
$$bias(Y_t|X,G) = |\mathbb{P}(Y_t = 0|G = 0) - \mathbb{P}(Y_t = 0|G = 1)|$$

Example (proxy predictor)

- $X \sim N(5 - G, \sqrt{5})$, $\mathbb{P}(G = 0) = \mathbb{P}(G = 1) = 0.5$
- $Y \sim Bernoulli(f(X)), f(x) = logistic(5 - x)$



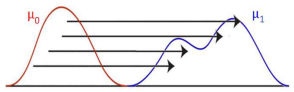




Global fairness metrics

Issue

- Threshold t in $\hat{Y}_t$ may not be known in advance; may be formulated in the form of an ML model.

❑ Wasserstein metric 

Cost of transporting the distribution of the protected class into non-protected:

$$Bias_{W_1}(f|X,G) := W_1(f(X)|G=0, f(X)|G=1)$$

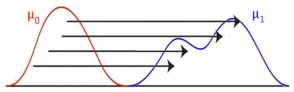
$$= \inf_{\pi \in \mathcal{P}(\mathbb{R}^2)} \left\{ \int |z_1 - z_2| \, \pi(dz_1, dz_2), \pi \text{ with marginals } P_{f(X)|G=k}, k \in \{0,1\} \right\}$$

$$= \int_0^1 |F_1(t) - F_0(t)| \, dt = \int_0^1 Bias^{st}\left(\hat{Y}_t | G\right) dt$$

Global fairness metrics

Issue

- Threshold t in $\hat{Y}_t$ may not be known in advance; may be formulated in the form of an ML model.

❑ Wasserstein metric 

Cost of transporting the distribution of the protected class into non-protected:

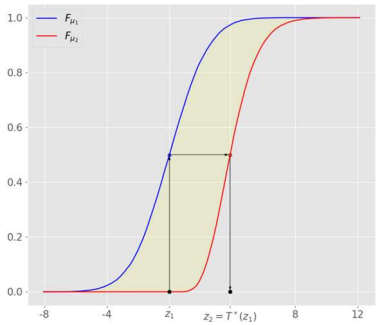
$$Bias_{W_1}(f|X,G) := W_1(f(X)|G=0, f(X)|G=1)$$

$$= \inf_{\pi \in \mathcal{P}(\mathbb{R}^2)} \left\{ \int |z_1 - z_2| \, \pi(dz_1, dz_2), \pi \text{ with marginals } P_{f(X)|G=k}, k \in \{0,1\} \right\}$$

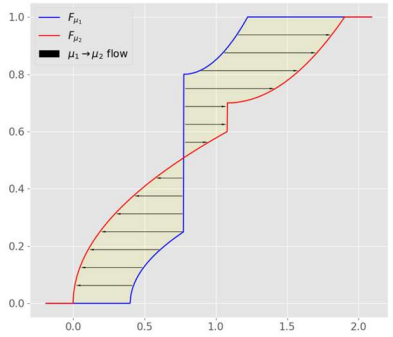
$$= \int_0^1 |F_1(t) - F_0(t)| \, dt = \int_0^1 Bias^{st}\left(\hat{Y}_t \middle| G\right) dt$$

❑ Generalizations: $\int_0^1 Bias\left(\hat{Y}_t \middle| G, \{\Omega_l\}_{l=1}^m\right) \cdot \mu\left(dt; \{F_{f|G=k,\Omega_l}\}\right)$

[Miroshnikov-Kotsiopoulos-Ravi Kannan (2022), Becker et al (2024)]



(a) Optimal transport map T^*.



(b) Transport flow of π^*.

2. Bias explanations

Individual feature attributions

(Local) Model explainer

Quantifies the contribution of an observation $x = (x_1, x_2, \dots x_n) \sim X$ to the value $f(x)$.

$$\mathbb{R}^n \ni x \to E(x; f, X, \mathcal{R}(f)) = (E_1, E_2, \dots E_n) \in \mathbb{R}^n \ .$$

Here the model f, the random vector X and model implementation $\mathcal{R}(f)$ serve as parameters.

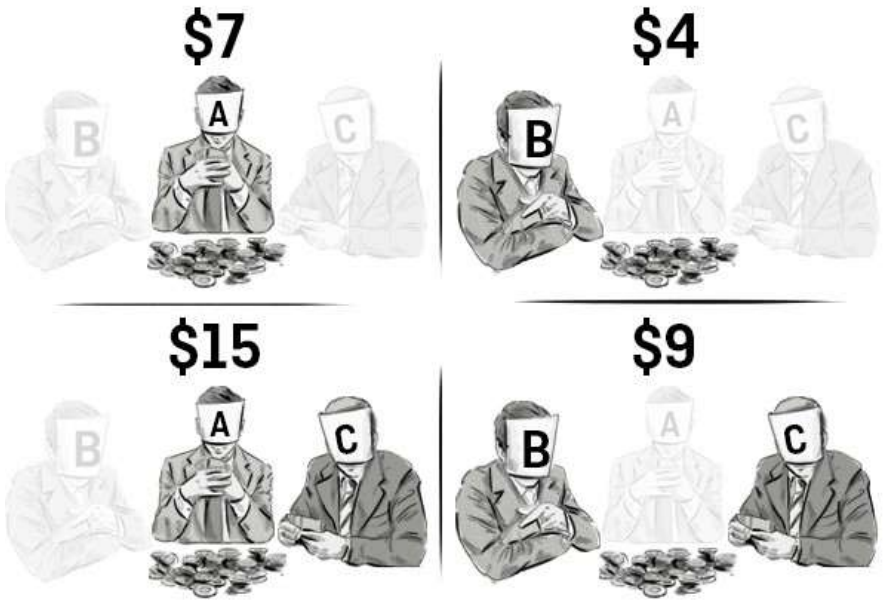
Game-theoretic approaches

- Game theoretic approaches have been explored in Štrumbelj & Kononenko (2014), Lundberg & Lee (2017)
- Cooperative game (N, v)
  - set of players indexed by $N = \{1,2, \dots, n\}$
  - utility $v(S)$, $S \subseteq N$
- Game value

$$(N, v) \to h[N, v] = \{h_i[N, v]\}_{i=1}^{n} \in \mathbb{R}^n$$

- Shapley value (Shapley, 1953)

$$\varphi_i[v] = \sum_{S \subseteq N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} \big(v(S \cup i) - v(S)\big),\ i \in N.$$



[from gametheory.online]

ML games and values

- We consider game values in the marginalist form

$$h_i[N, v] = \sum_{S \subseteq N \setminus \{i\}} w(S, |N|) \cdot \big(v(S \cup i) - v(S)\big)$$

- ML games
  - $v^{CE}(S, x; X, f) = \mathbb{E}[f(X_S, X_{-S})|X_S = x_S]$ (conditional game)
  - $v^{ME}(S, x; X, f) = \mathbb{E}[f(x_S, X_{-S})]$ (marginal game)

- Linearity

  $f \to h[N, v(\cdot, x; X, f)]$ is linear in models f.

ML Fairness Explainability

A.M., K. Kotsiopoulos, R. Franks, A. Ravi Kannan, "Wasserstein-based fairness interpretability framework for machine learning models, Machine Learning (Springer), 2022.

- Global metric

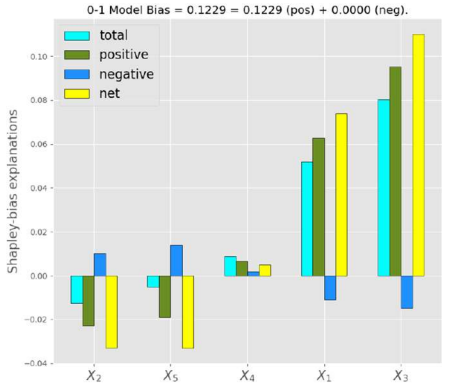
$$\text{Bias}_{W_1}(f|G) = W_1\left(P_{f(X)|G=0}, P_{f(X)|G=1}\right) = \int_0^1 bias_t^C\,(f|G)dt$$

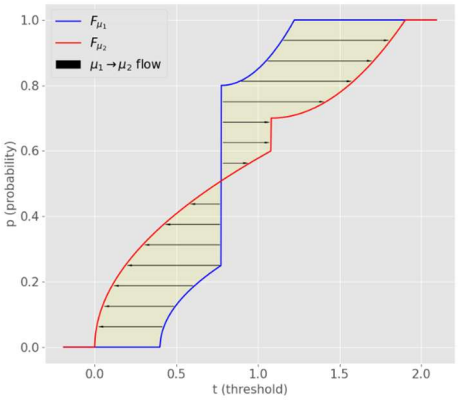
- Marginal bias game

$$v_{bias}(S;f) = Bias_{W_1}(v^{ME}(S;f,X)|X,G)$$

- Bias explanations

$$\varphi_i[v_{bias}(S;f)], i \in N = \{1,2,\dots N\}$$





2. Bias mitigation

Mitigation under regulatory constraints

Demographically blind models

- ❑ Use of G during ML training is typically not allowed.
- ❑ Use of G at inference time is not allowed, i.e. f must explicitly depend on X only.
- ❑ Threshold t may be not known in advance. Global bias metrics must be used.

e.g. [Hardt et al (2015), Feldman et al (2015), Gordaliza et al (2019), Kwegyir-Aggrey et al (2023)]

Use of G in validation and mitigation (after ML training)

- ❑ Proxy for G are used for bias measurement by the compliance office after ML training.
- ❑ Bias mitigation for the trained ML model is performed by the compliance office.

Problem setup

Given a parametrized family of models $\mathcal{C} = \{f(x;\theta): \theta \in \mathbb{R}^m\}$ consider the minimization

(OP) $$f(x;\theta^*(\omega)) = \underset{f\in\mathcal{C}}{\operatorname{argmin}} \{ \mathbb{E}[\mathcal{L}(f(X;\theta),Y)] + \omega \cdot \mathcal{B}(f(\cdot;\theta)|\mathrm{X},G) \}$$

where $\omega \in (0,\infty)$ is a penalization parameter and $\mathcal{B}(f,X|G)$ is a global fairness metric.

Problem setup

Given a parametrized family of models $\mathcal{C} = \{f(x;\theta): \theta \in \mathbb{R}^m\}$ consider the minimization

(OP) $$f(x;\theta^*(\omega)) = \underset{f\in\mathcal{C}}{\operatorname{argmin}} \{ \mathbb{E}[\mathcal{L}(f(X;\theta),Y)] + \omega \cdot \mathcal{B}(f(\cdot;\theta)|\mathrm{X},G) \}$$

where $\omega \in (0,\infty)$ is a penalization parameter and $\mathcal{B}(f,X|G)$ is a global fairness metric.

- Bias metrics of interest: [Jiang et al 2020, Vogel et al 2022, Becker et al 2024, Franks-Miroshnikov 2025]
  - $Bias_{W_1}(f|X,G) \coloneqq \int \left|F^{[-1]}_{f|G=0}(q) - F^{[-1]}_{f|G=1}(q)\right| dq,\ p \geq 1$
  - $Bias_{W_1,\mu}(f|X,G) \coloneqq \int_0^1 \left|F_{f|G=0}(t) - F_{f|G=1}(t)\right| \mu(dt)$
  - $Bias^{(c)}_{\mu}(f|X,G) \coloneqq \int_0^1 c\left(F_{f|G=0}(t), F_{f|G=1}(t)\right) \mu(dt),\ \ cost = c(\cdot,\cdot) \geq 0$
- Note: care must be taken to to compute $\nabla_\theta\, Bias(f(\cdot;\theta)|\, X,G)$

Optimization approaches

- **(A1)** Optimization during ML training. $\mathcal{C}$ is the family of ML models where θ is ML parameter.

[Jiang et al 2020, Vogel et al 2022].

- **(A2)** $\mathcal{C}$ is the family of ML models where θ is a hyper parameter. [Perrone 2021]

- **(A3)** Optimize after ML training f_*. Then $\mathcal{C} = \mathcal{C}(f_*)$ is the family of postprocessed models.

[Miroshnikov-Kotsiopoulos-Franks-Ravi Kannan (2021)], [US Patent 12002258 B2 (2024)]

[Franks-Miroshnikov arXiv:2504.01223 (2025)]

Notes

- (A1)-(A2) can be computationally costly, (A1) may not be feasible (e.g. tree-based models).
- (A3) $\mathcal{C}(f_*)$ has to be carefully formulated not to impact performance and $f(x;\theta)$ must be explainable.

Choice of a family

Setup: Given a family of models $\mathcal{C} = \{f(x;\theta): \theta \in \mathbb{R}^m\}$ consider the minimization problem

(OP) $$f^{(\omega)} = argmin_{\mathcal{C}}\left(\mathbb{E}[\mathcal{L}(f(X;\theta),Y)] + \omega \cdot \mathcal{B}(f,X|G)\right), \ \omega \in (0,\infty),$$

where ω is a penalization parameter and $\mathcal{B}(f,X|G)$ is a global fairness metric (e.g. W_1).

Approaches

(A1) Optimize during ML training: $\mathcal{C}$ is the family of models, e.g. [Vogel et al 2022].

(A2) Optimize after ML training f_*. Then $\mathcal{C} = \mathcal{C}(f_*)$ is the family of postprocessed models.

Practical challenges

- (A1) is computationally costly and maybe difficult for some classes $\mathcal{C}$ of ML models (e.g. tree-based models)
- (A2) $\mathcal{C}(f_*)$ has to be carefully formulated not to impact performance too much.
- $f^{(\omega)}$ must be explainable.

3. Post processing

Postprocessing family 1 (perturbing input)

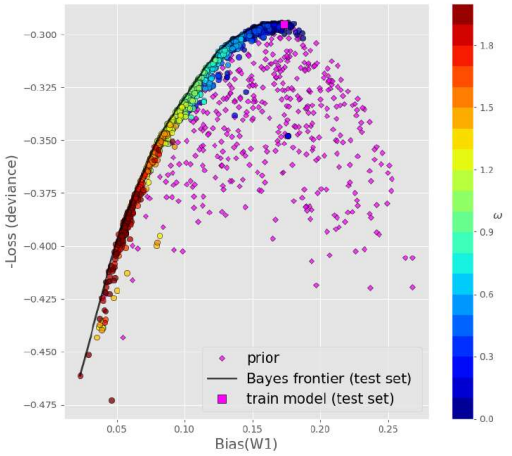
- Given a trained ML model $f_*(x)$, define a scaling family about f_*

$$\mathcal{C}(f_*) = \{f(x,\theta)\colon T_0 \circ f_*(T_1(x_1), T_2(x_2), \dots, T_n(x_n))\}$$

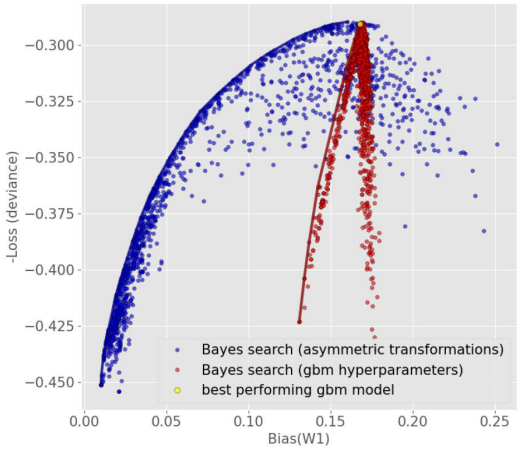
where $T_i(t) = \theta_i^{(1)} \cdot t + \theta_i^{(2)}$ or some variations of that.

[Miroshnikov-Kotsiopoulos-Franks-Ravi Kannan (2021) & US Patent 12002258 B2 (2024)]

- Use Bayesian optimization or SGD to solve (OP) over $\mathcal{C}$ to get the Pareto frontier.



Issues
- Bayesian optimization is slow in high dimensions.
- SGD is not always appropriate.



- To reduce dimension, determine features most contributing to bias:
  - Game $v_{bias}(S;f) = Bias_{W_1}(v^{ME}(S;f,X)|X,G),\ S \subseteq N.$
  - Shapley values $\varphi_i[v_{bias}(S;f)],\ i \in N.$
  - Set $T_i(t) = t$ for non-contributing $i \in N.$

[Miroshnikov-Kotsiopoulos-Franks-Ravi Kannan, ML Springer (2022)]

Postprocessing family 2 (perturbing input)

[Miroshnikov-Kotsiopoulos-Ravi Kannan-Dickerson-Franks US 20220414766 Patent Application]
& [Franks-Miroshnikov arXiv:2504.01223 (2025)]

❑ Given a trained model f_*, define

$$\mathcal{C}(f_*) = \left\{ f(x,\theta):\ f(x,\theta) = f_*(x) - \sum_{j=0}^{m} \theta_j \cdot w_j(x; f_*, X)\,,\ \theta \in \Theta \subseteq \mathbb{R}^{m+1} \right\}$$

where $w_j(x; f_*, X)$ are weight functions (or encoders) that depend on f_* and X, with $w_0 \equiv 1$.

❑ Encoders $w(f_*, X)$:

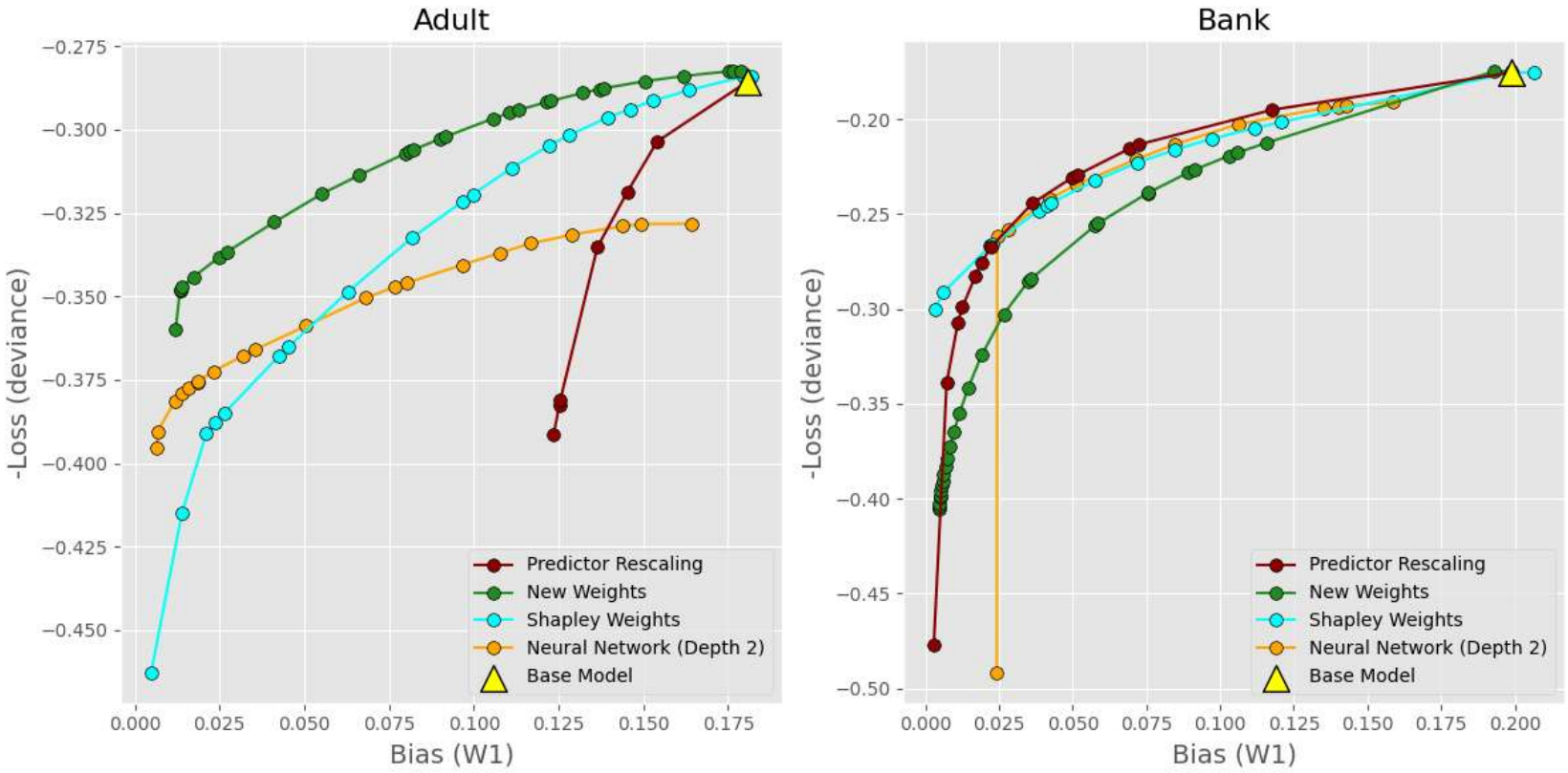
- Features or additive models of features (e.g., biased features X_i).
- Explanations (e.g., marginal Shapley values $\varphi^{ME}(x; f_*, X)$).
- Model representation components (i.e., trees or their linear combinations when f_* is a tree ensemble).

❑ Properties

- The problem is linear in w.
- If f_* and w_j are explainable, then $f^{(\omega)}$ is explainable (assuming the explanation map is linear).

Public dataset

- Adult dataset (12 variables, $G = Gender$, 48842 samples)
- Bank dataset (19 variables, $G = Age$, 41188 samples)



[Franks-Miroshnikov arXiv:2504.01223 (2025)]