Measuring proxy discrimination through model distortions

Andreas Tsanakas
Bayes Business School

joint work with
M. Lindholm, S. Pesenti, R. Zhu

Insurance Data Science Conference
Hannover, June 2026

Outline

Setup

Y – claims costs

$\mathbf{X}$ – **permitted covariates**

D – **sensitive attribute**, taking values in $\mathfrak{D} = \{d_0, \ldots, d_k\}$

$\mu(\mathbf{X}, D) = \mathbb{E}[Y \mid \mathbf{X}, D]$ – **best-estimate price**

- Direct use of D typically not allowed [Frees and Huang, 2023]

$\mu(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ – **unawareness price**

- Basis for technical prices in practice

Example: Synthetic motor dataset [So et al., 2021]

100,000 policy records with claims and conventional + telematics covariates.

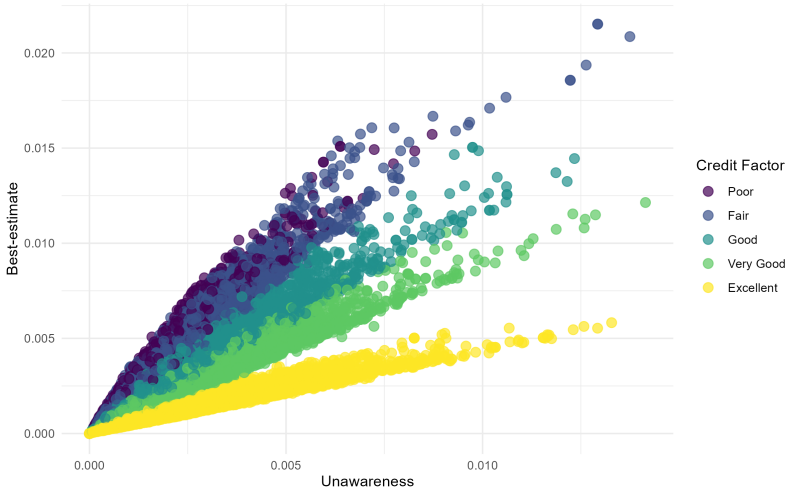
Sensitive attribute: credit score, discretised into 5 ordered bands

Category	Poor	Fair	Good	Very Good	Excellent
Frequency	0.026	0.059	0.107	0.170	0.638

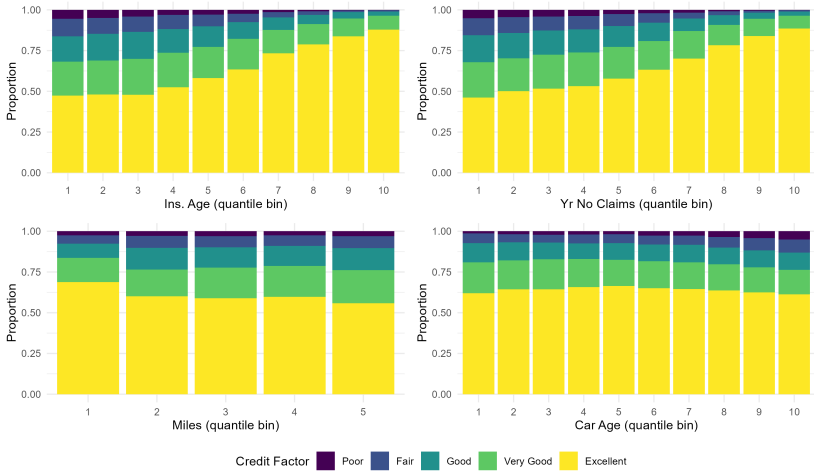
Models:

- Classifier $\mathbb{P}(D \mid \mathbf{X})$: ordered logistic regression
- Claim amount $\mu(\mathbf{X}, D)$: Tweedie GAM with log-link and exposure offset
- Covariates: insured age, car age, miles, no-claims years, marital status, car use, territory

Best-estimate $\mu(\mathbf{X}, D)$ vs unawareness prices $\mu(\mathbf{X})$



Dependence of credit score D on covariates $\mathbf{X}$



Proxy discrimination

The unawareness price decomposes as:

$$\mu(\mathbf{X}) = \sum_{j=0}^{k} \mu(\mathbf{X}, d_j)\, \mathbb{P}(D = d_j \mid \mathbf{X})$$

- The classifier $\mathbb{P}(D \mid \mathbf{X})$ **implicitly infers** D from $\mathbf{X}$
- Selection or causal effects [Côté et al., 2025]

  $\rightarrow$ **proxy discrimination**: prices remain sensitive to D via the classifier

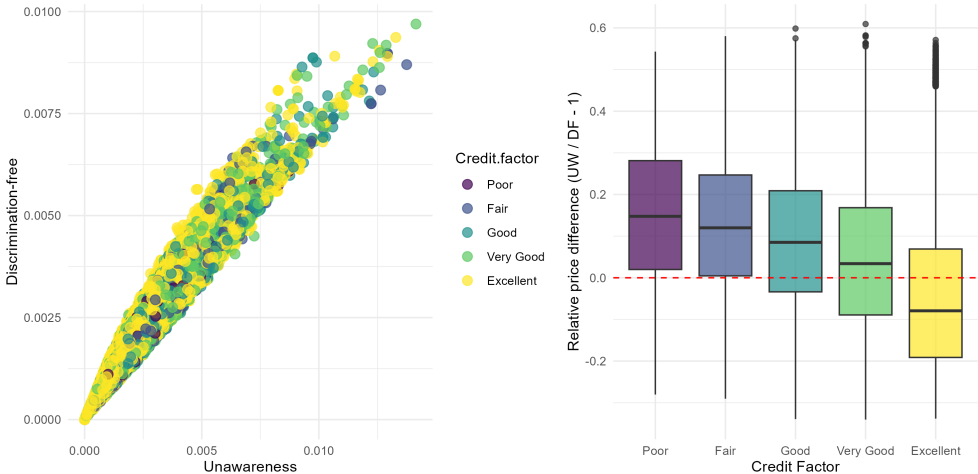
Removing proxy discrimination

The **discrimination-free price** as proposed by [Pope and Sydnor, 2011, Lindholm et al., 2022] is

$$\mu^*(\mathbf{X}) = \sum_{j=0}^{k} \mu(\mathbf{X}, d_j)\, \mathbb{P}(D = d_j)$$

- Replace classifier by the (an) **unconditional distribution** of D

Discrimination-free vs unawareness prices



Outline

Materiality and sensitivity

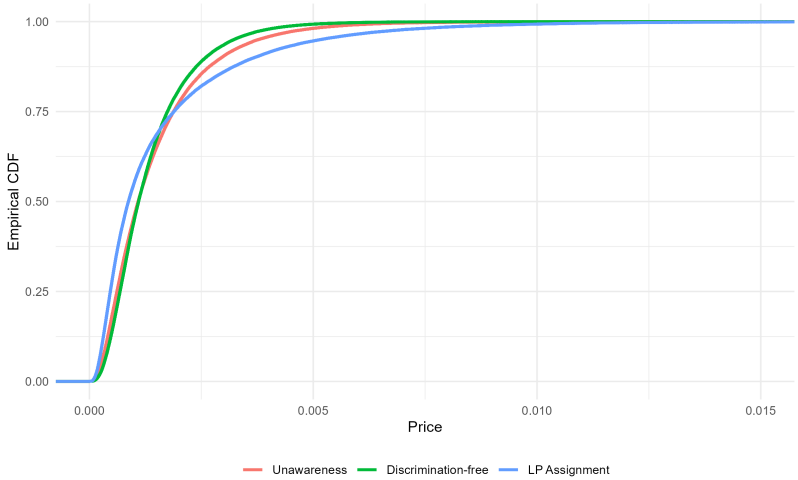
Proxy discrimination will generally be present in portfolios, but **how material is the effect**?

Discriminatory effects arise because of the implicit **sensitivity of prices to the classifier** $\mathbb{P}(D \mid \mathbf{X})$ [Lindholm et al., 2024, Steensgaard et al., 2026]

Key idea: **distort the classifier and monitor impact on prices**

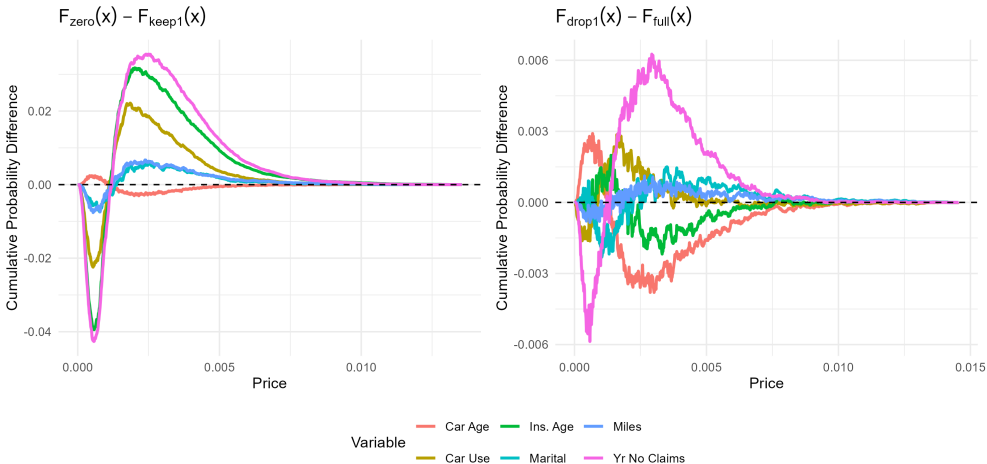
- Discrimination-free price: distort $\mathbb{P}$ so that D and $\mathbf{X}$ become **independent**, while maintaining cost structure
- Alternatively: **increase dependence** by assigning conditional probabilities in $\{0, 1\}$ (a transportation problem)

Comparison of price distributions (non-mean preserving)



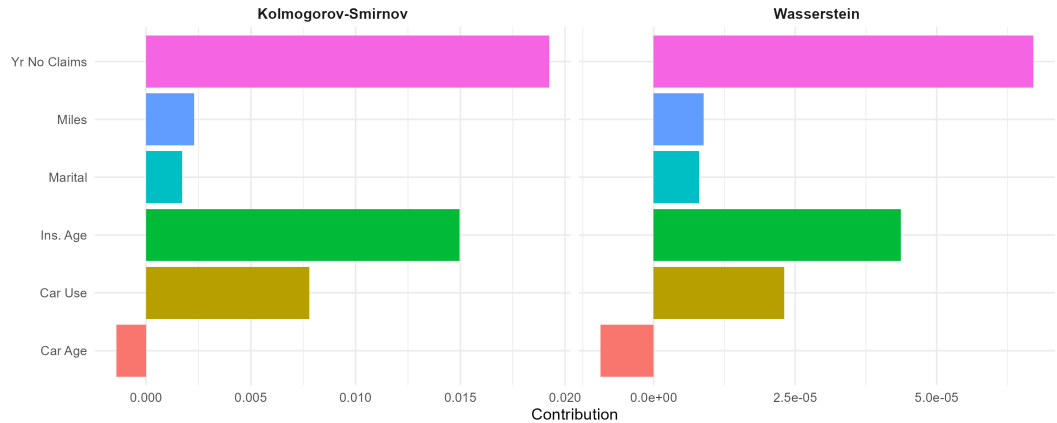
Drivers of proxy discrimination

Refit classifier on covariate subsets



Shapley attributions

Allocate the distance between discrimination-free and unawareness price distributions



Outline

How to choose a distortion?

For a distorted classifier $\mathbb{Q}(D \mid \mathbf{X})$ the resulting unawareness price becomes:

$$\mu(\mathbf{X}; \mathbb{Q}) = \sum_{j=0}^{k} \mu(\mathbf{X}, d_j)\, \mathbb{Q}(D = d_j \mid \mathbf{X})$$

Convex order: look for bounds $\mathbb{Q}^-, \mathbb{Q}^+$ such that for any convex function v

$$\mathbb{E}\left[v(\mu(\mathbf{X}; \mathbb{Q}^-))\right] \leq \mathbb{E}\left[v(\mu(\mathbf{X}))\right] \leq \mathbb{E}\left[v(\mu(\mathbf{X}; \mathbb{Q}^+))\right]$$

$\rightarrow$ More/less spread out prices with the same mean

Lower convex bound: a smoothed classifier

Smooth $\mathbb{P}(D \mid \mathbf{X})$ by averaging probabilities across policyholders with similar best-estimate price scenarios

$$\mathbb{Q}^{-}(D = d_j \mid \mathbf{X}) = \mathbb{E}\Big[\mathbb{P}(D = d_j \mid \mathbf{X}) \mid \mu(\mathbf{X}, d_0), \ldots, \mu(\mathbf{X}, d_k)\Big]$$

Not the same as the discrimination-free price:

- Preserves the mean (discrimination-free prices don't)
- Guarantees less dispersed prices (discrimination-free prices don't)
- A **different notion of non-discrimination**: measurability wrt best-estimate scenarios $\mu(\mathbf{X}, d_0), \ldots, \mu(\mathbf{X}, d_k)$

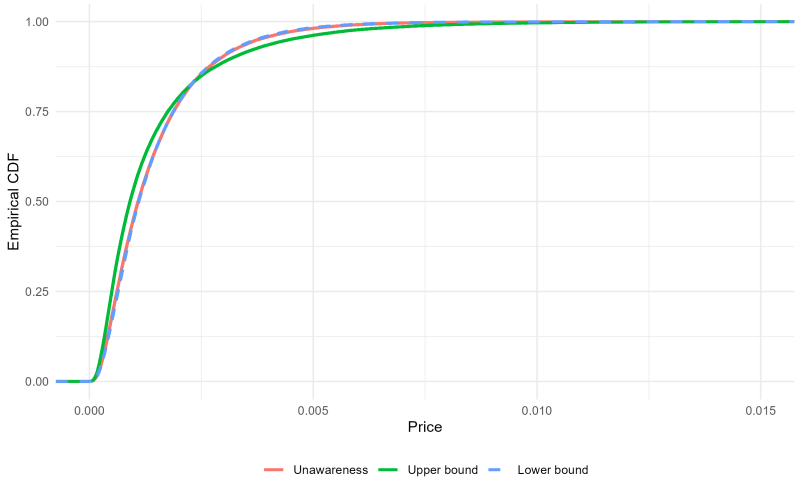
Upper convex bound: a randomised sharp classifier

We now add fictitious information by **randomisation**

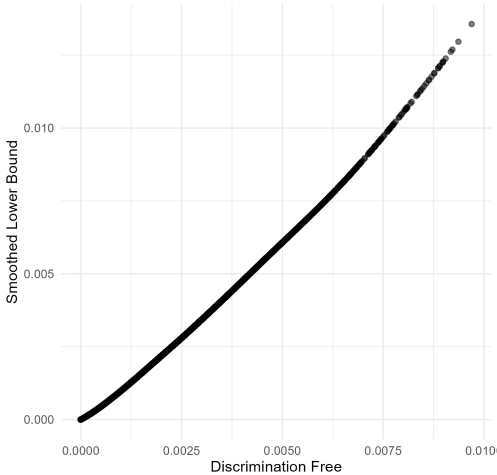
- Each policyholder is assigned to a single category of D with **probability 1**
- The assigned category is drawn at random, with probabilities $\mathbb{P}(D = d_j \mid \mathbf{X} = \mathbf{x}_i)$

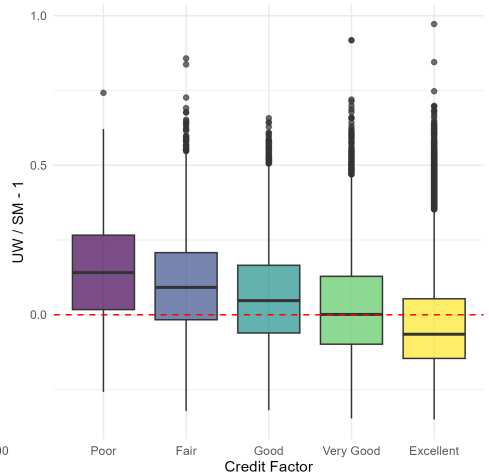
This is equivalent to adding a **pure proxy variable** Z to the model, which allows perfect prediction of D, without impacting claims cost predictions

Comparison of price distributions (mean preserving)

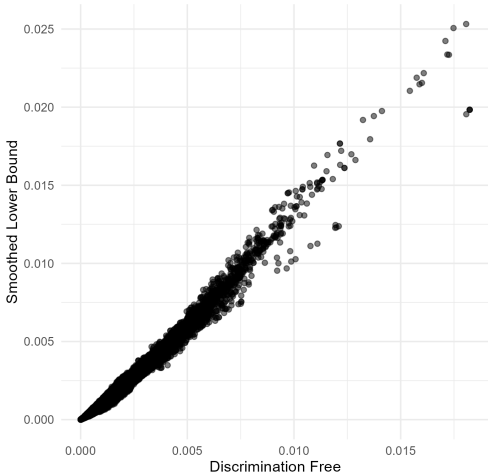


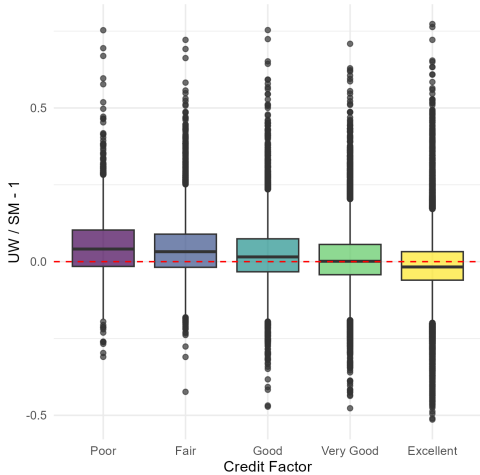
What does the lower bound achieve?





What does the lower bound achieve? (XGBoost version)





Conclusions

Distorting predictions of sensitive features

- Materiality via change in price distribution
- Sharper classifiers amplify proxy effects and produce more dispersed prices

Lower bound and proxy discrimination

- The convex lower bound is obtained by regressing the predictions of D on best-estimate scenarios $\mu(\mathbf{X}, d_j)$
- Mean-preserving alternative to [Lindholm et al., 2022]
- A more general principle: *prices must be functions only of costs under different D-scenarios*

Côté, O., Côté, M.-P., and Charpentier, A. (2025).
A fair price to pay: Exploiting causal graphs for fairness in insurance.
Journal of Risk and Insurance, 92(1):33–75.

Frees, E. W. and Huang, F. (2023).
The discriminating (pricing) actuary.
North American Actuarial Journal, 27(1):2–24.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2022).
Discrimination-free insurance pricing.
ASTIN Bulletin, 52:55–89.

Lindholm, M., Richman, R., Tsanakas, A., and Wuthrich, M. V. (2024).
What is fair? proxy discrimination vs. demographic disparities in insurance pricing.

Scandinavian Actuarial Journal, 2024(9):935–970.

Pope, D. G. and Sydnor, J. R. (2011).
Implementing scrimination policies in statistical profiling models.
American Economic Journal: Economic Policy, 3(3):206–31.

So, B., Boucher, J.-P., and Valdez, E. A. (2021).
Synthetic dataset generation of driver telematics.
Risks, 9(4):58.

Steensgaard, T., Hiabu, M., and Pfister, N. (2026).
Invariance in the presence of protected features.
Presented at the 1st ASTIN Bulletin Conference, Zurich, January 2026.