Fairness, Variance and Stability

Falaah Arif Khan
New York University
https://falaaharifkhan.github.io/research/
Twitter: @FalaahArifKhan Email: fa2161@nyu.edu

On Fairness and Stability

Preprint (work in progress): https://arxiv.org/abs/2302.04525

Motivation

Fairness Metrics — composed of group-specific error metrics*

Equal opportunity = True Positive Rate (dis) - True Positive Rate (priv)

Statistical Parity Difference = Positive Rate (dis) - Positive Rate (priv)

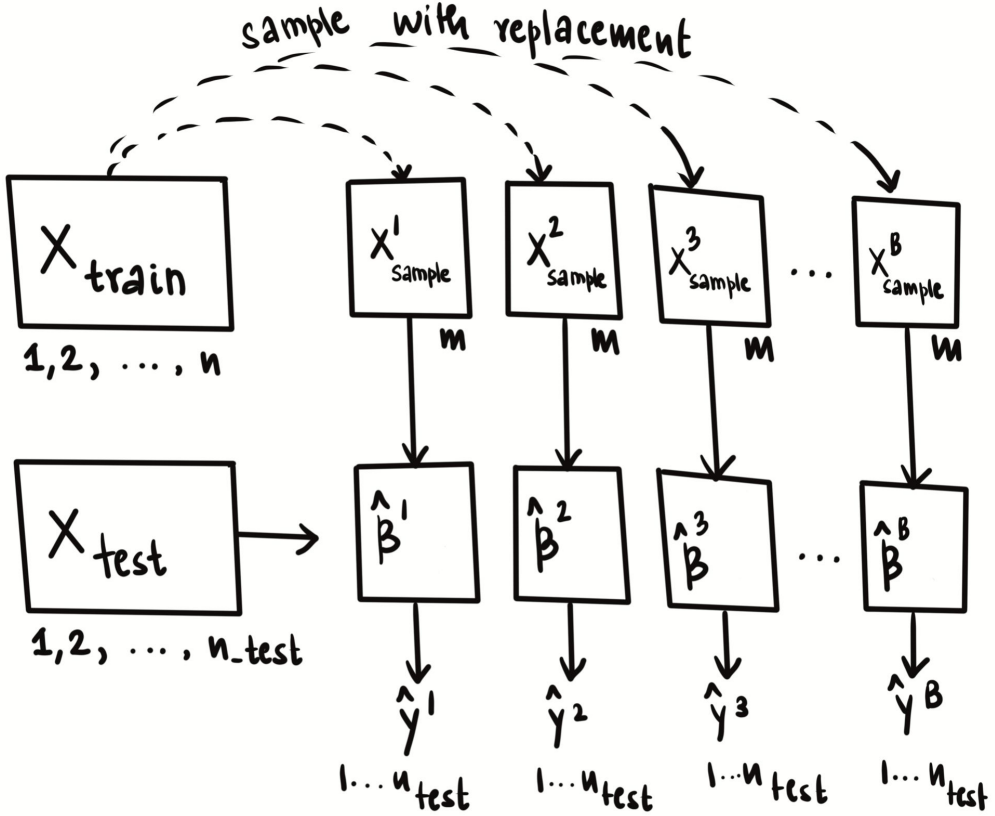
Disparate Impact = Positive Rate (dis) / Positive Rate (priv)

Accuracy Parity = Accuracy (dis) - Accuracy (priv)

*Ratio or difference between a base measure computed on priv and dis groups

Estimating variance – “Sampling during Inference”

The Bootstrap



(Efron, 1979)

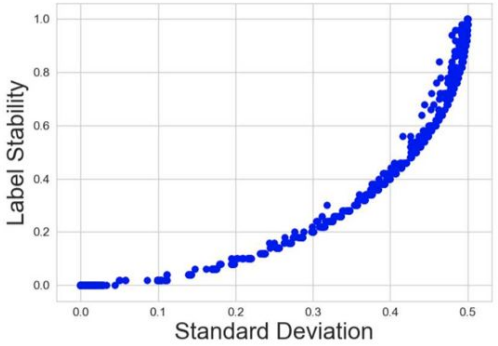
Variance Metrics

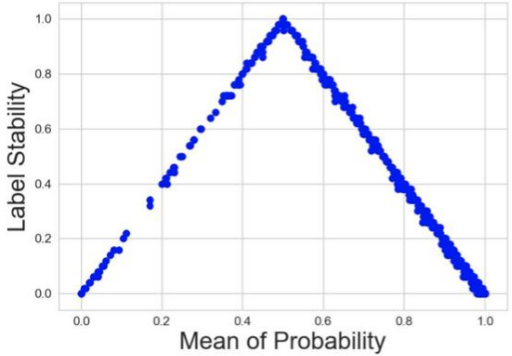
Standard deviation [1]

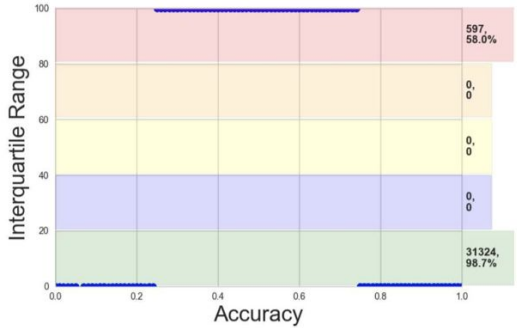
Inter-quantile range [1]

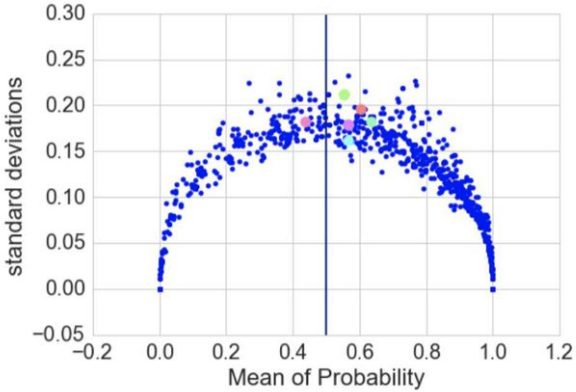
Label stability [1]

Jitter [2]









[1] Towards Uncertainty Quantification for Supervised Classification (Darling et al, 2018)

[2] Model Stability with Continuous Data Updates (Liu et al, 2022)

Variance Metrics

Standard deviation [1]

Inter-quantile range [1]

Label stability [1]

Jitter [2]

2.2.1 Label Stability. Label Stability [8] is defined as the normalized absolute difference between the number of times a sample is classified as positive or negative:

$$\text{Label Stability} = \frac{|\sum_{i=1}^{b} \mathbb{1}[p_{\theta_i}(x) == 1] - \sum_{i=1}^{b} \mathbb{1}[p_{\theta_i}(x) == 0]|}{b}$$

where x is an unseen test sample, and $p_{\theta_i}(x)$ is the prediction of the i^{th} model in the ensemble that has b estimators.

2.2.2 Jitter. Jitter [21] is a measure of the disparities of the model's predictions for each individual test example. It reuses a notion of *Churn* [24] to define a "pairwise jitter":

$$J_{i,j}(p_\theta) = \text{Churn}_{i,j}(p_\theta) = \frac{\left|p_{\theta i}(x) \neq p_{\theta j}(x)\right|_{x \in X}}{|X|}$$

where x is an unseen test sample, and $p_{\theta i}(x)$, $p_{\theta j}(x)$ are the predictions of the i^{th} and j^{th} estimator in the ensemble for x, respectively.

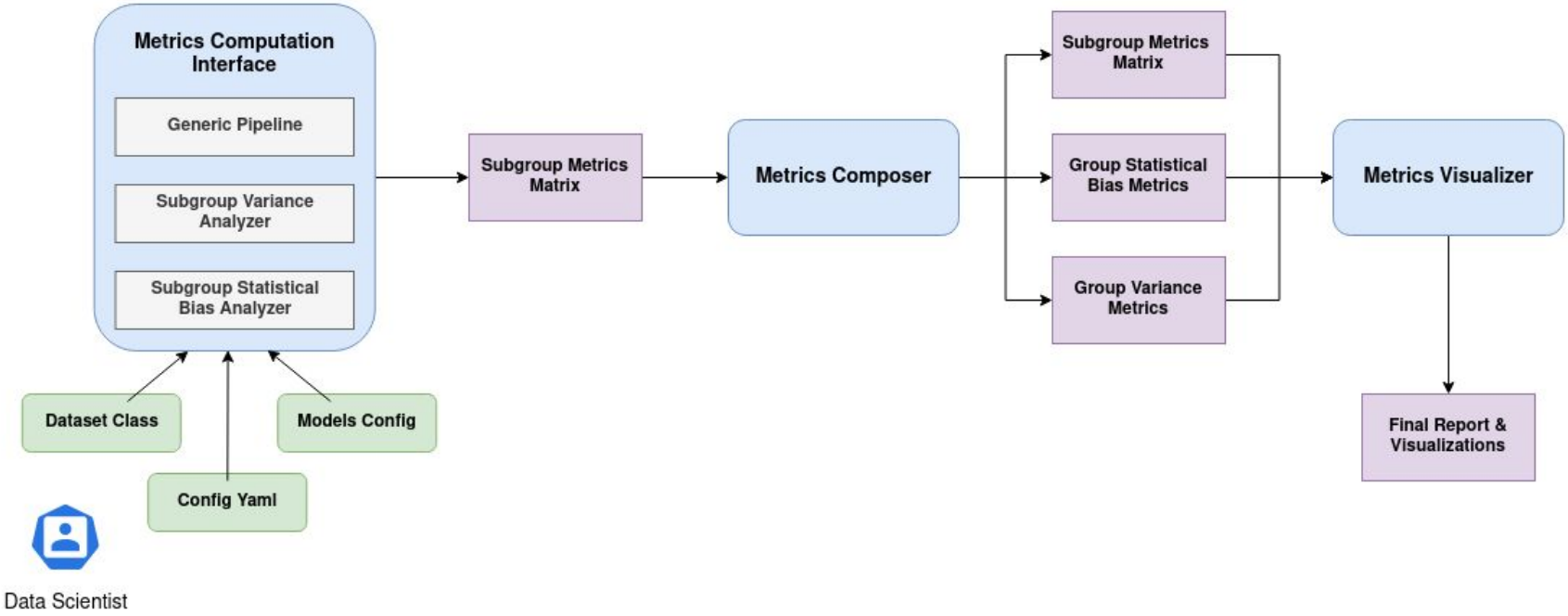
To compute the variability over all models in the ensemble, we need to average *pairwise jitters* over all pairs of models. This more general definition is called *Jitter*:

$$J(p_\theta) = \frac{\sum_{\forall i,j \in N} J_{i,j}(p_\theta)}{N \cdot (N-1) \cdot \frac{1}{2}}, \text{ where } i < j$$

[1] Towards Uncertainty Quantification for Supervised Classification (Darling et al, 2018)

[2] Model Stability with Continuous Data Updates (Liu et al, 2022)

The Virny software library



https://github.com/DataResponsibly/Virny

Reconciling error-based and variance-based analysis

(folktables)

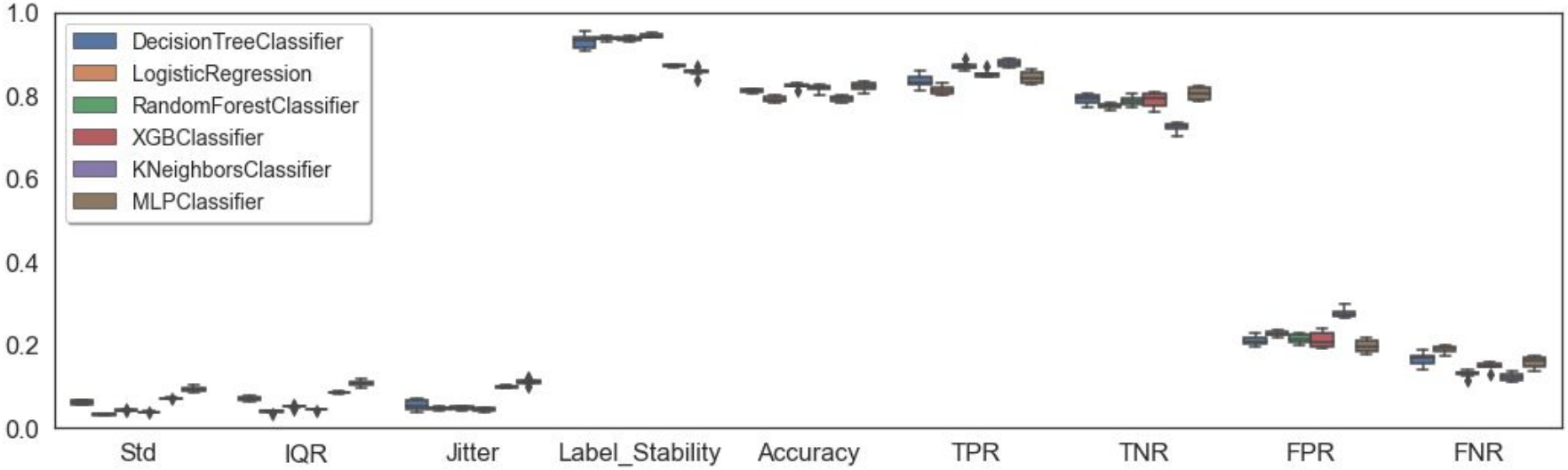


Table 3. Statistical bias-based and variance-based fairness metrics on **folktables**, averaged over 6 runs. RF is Random Forest, DT is Decision Tree, XGB is XGBoost, MLP is Multi-Layer Perceptron, LR is Logistic Regression and kNN is k-Nearest Neighbors

	RF	DT	XGB	MLP	LR	kNN
Accuracy Parity (sex)	-0.0793	-0.0662	-0.0656	-0.0634	-0.0618	-0.0588
Accuracy Parity (race)	-0.0057	0.0076	-0.0013	0.0015	-0.0010	0.0069
Accuracy Parity (sex&race)	-0.0756	-0.0545	-0.0597	-0.0560	-0.0587	-0.0508
Equalized Odds FPR (sex)	0.1047	0.0991	0.0592	0.0487	0.0124	0.0244
Equalized Odds FPR (race)	0.0238	0.0152	0.0135	-0.0154	-0.0069	-0.0086
Equalized Odds FPR (sex&race)	0.1278	0.1176	0.0716	0.0323	0.0056	0.0246
Statistical Parity Difference (sex)	0.1450	0.1678	0.0474	0.0293	-0.0328	0.0086
Statistical Parity Difference (race)	0.1048	0.0992	0.0596	0.0105	0.0246	0.0644
Statistical Parity Difference (sex&race)	0.1998	0.2316	0.0785	0.0205	-0.0268	0.0682
Disparate Impact (sex)	1.1351	1.1644	1.0436	1.0274	0.9705	1.0071
Disparate Impact (race)	1.0942	1.0928	1.0546	1.0097	1.0227	1.0536
Disparate Impact (sex&race)	1.1944	1.2350	1.0746	1.0195	0.9753	1.0568
IQR Parity (sex)	0.0027	-0.0045	0.0029	0.0210	0.0017	0.0110
IQR Parity (race)	-0.0016	-0.0032	0.0037	0.0109	0.0068	-0.0066
IQR Parity (sex&race)	0.0021	-0.0065	0.0072	0.0312	0.0091	0.0044
Jitter Parity (sex)	0.0153	0.0242	0.0126	0.0374	0.0035	0.0418
Jitter Parity (race)	0.0017	0.0083	0.0085	0.0158	0.0034	-0.0093
Jitter Parity (sex&race)	0.0168	0.0335	0.0213	0.0498	0.0085	0.0272
Std Parity (sex)	0.0016	-0.0028	0.0026	0.0165	0.0013	0.0091
Std Parity (race)	-0.0013	0.0008	0.0034	0.0086	0.0051	-0.0046
Std Parity (sex&race)	0.0009	-0.0014	0.0065	0.0246	0.0069	0.0047
Label Stability Ratio (sex)	0.9770	0.9641	0.9810	0.9448	0.9964	0.9394
Label Stability Ratio (race)	0.9977	0.9884	0.9880	0.9746	0.9952	1.0141
Label Stability Ratio (sex&race)	0.9754	0.9512	0.9688	0.9262	0.9892	0.9609

Revisiting Bias-Variance Trade-offs in the Context of Fair Prediction

Preprint (work in progress): https://arxiv.org/abs/2302.08704

Motivation:

1. Decomposition of model error [1]
   Error = Statistical bias + Variance + Noise

2. Randomness is neutral — variance could be more morally acceptable than statistical bias.

Idea: Can disparity in variance across groups be exploited to design fairness-enhancing interventions?

[1] Domingos, P.M. A Unified Bias-Variance Decomposition.

Conditional-IID

$$\mathcal{D}^{priv} = \{(X^i, Y^i) | X^i_{protected} = x^*\}, i = 1, 2 \ldots n$$

$$\mathcal{D}^{dis} = \{(X^i, Y_i) | X^i_{protected} \neq x^*\}, i = 1, 2 \ldots n$$

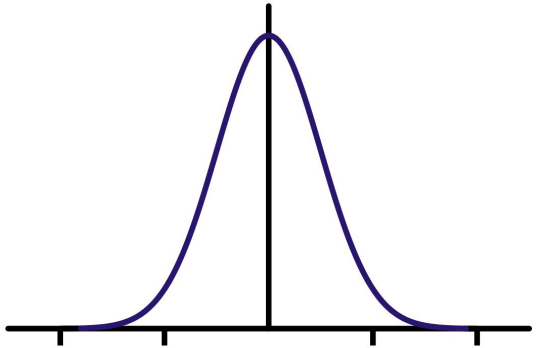
$$\mathcal{D} = \mathcal{D}^{priv} \cup \mathcal{D}^{dis}$$

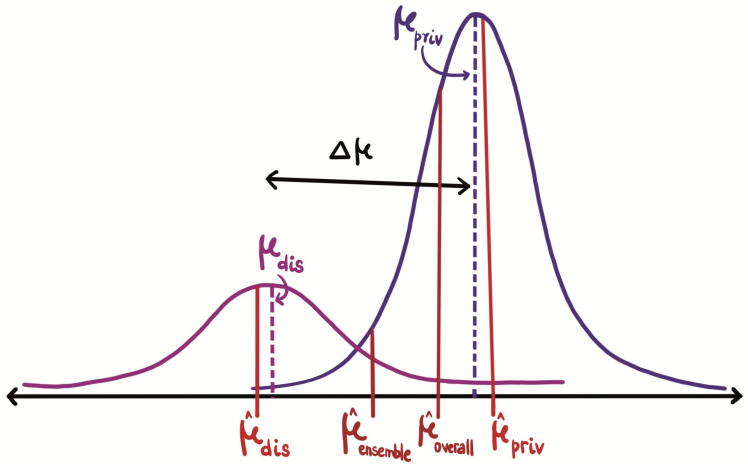
Conventionally, we assume that samples (X_i, Y_i) are i.i.d

$$\mathcal{D}^{priv}, \mathcal{D}^{dis} \sim (\mathcal{X}, \mathcal{Y})$$

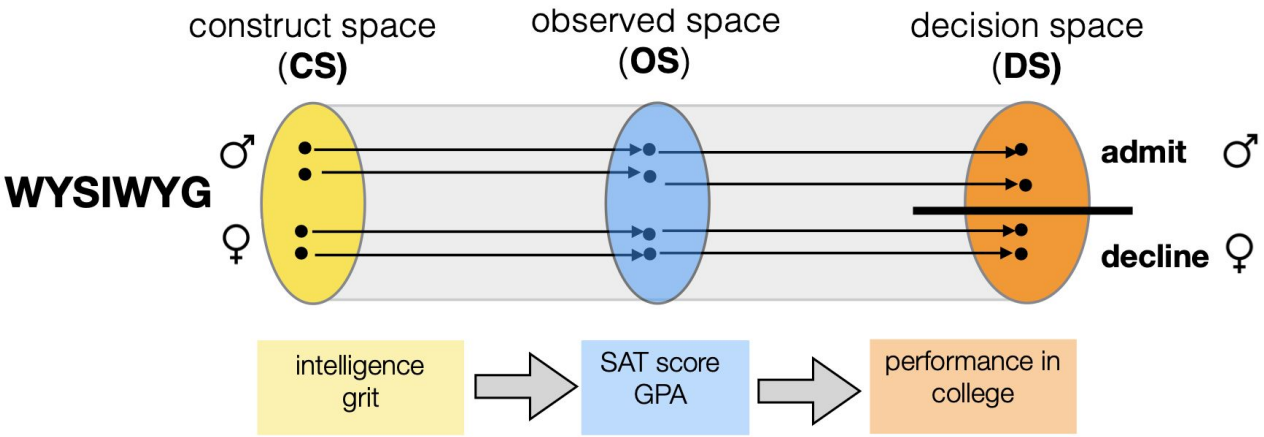
Instead, in this paper we model the conditional-i.i.d. setting

$$\mathcal{D}^{priv} \sim (\mathcal{X}^{priv}, \mathcal{Y}^{priv}), \mathcal{D}^{dis} \sim (\mathcal{X}^{dis}, \mathcal{Y}^{dis})$$





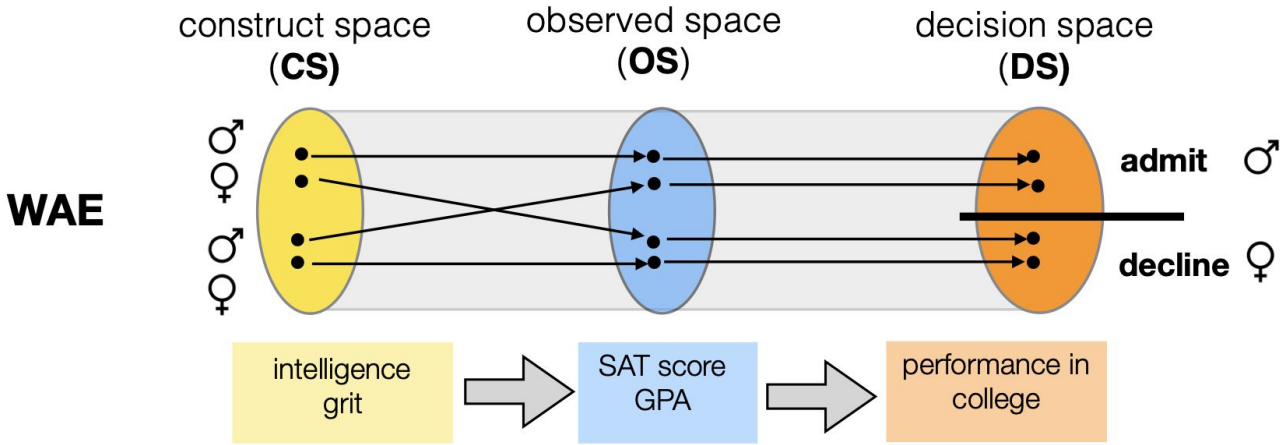
IID and WYSIWYG



What you see is what you get (**WYSIWYG**): there exists a mapping from CS to OS that has low distortion. That is, we believe that OS faithfully represents CS. **This is the individual fairness world view.**

[1] Friedler et al. **On the (im)possibility of fairness**

Conditional-IID and WAE



We are all equal (**WAE**): the mapping from **CS** to **OS** introduces **structural bias** - there is a distortion that aligns with the group structure of **CS**. **This is the group fairness world view.**

[1] Friedler et al. **On the (im)possibility of fairness**

(iid)

$$\hat{f}_{overall}(\mathcal{D}) := argmin_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell(f(X), Y)] \tag{7}$$

$$\hat{y}_{iid}(X) = \hat{f}_{overall}(X_{relevant}, X_{protected}) \tag{8}$$

(conditional-iid)

$$\hat{f}_{priv}(\mathcal{D}) := argmin_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell(f(X^{priv}), Y^{priv})] \tag{9}$$

$$\hat{f}_{dis}(\mathcal{D}) := argmin_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}[\ell(f(X^{dis}), Y^{dis})] \tag{10}$$

$$\begin{aligned} \hat{y}_{ciid}(X) = \hat{f}_{priv}(X_{relevant}).\mathbb{1}[X_{protected} = x^*] \\ +\hat{f}_{dis}(X_{relevant}).\mathbb{1}[X_{protected} \neq x^*] \end{aligned} \tag{11}$$

We will also look at these models in isolation, i.e., if we applied a single conditional model to the entire population:

$$\hat{y}_{priv}(X) = \hat{f}_{priv}(X_{relevant}) \tag{12}$$

$$\hat{y}_{dis}(X) = \hat{f}_{dis}(X_{relevant}) \tag{13}$$

Theoretical Analysis

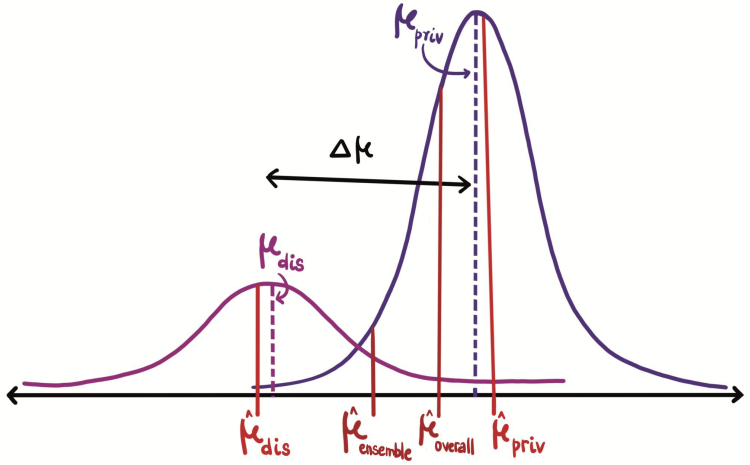


Table 1. Summary of bias-variance trade-offs of different mean estimators

Model	Bias on $priv$	Bias on dis	Variance
overall (i.i.d.)	$p_{dis}.\mathbb{E}[\Delta\mu]$	$p_{priv}.\mathbb{E}[\Delta\mu]$	$\frac{1}{n}[\frac{n_{priv}}{n}.\sigma^2_{priv} + \frac{n_{dis}}{n}.\sigma^2_{dis}]$
ensemble	$\frac{1}{2}\mathbb{E}[\Delta\mu]$	$\frac{1}{2}\mathbb{E}[\Delta\mu]$	$\frac{1}{4}(\frac{\sigma^2_{priv}}{n_{priv}} + \frac{\sigma^2_{dis}}{n_{dis}})$
disprivileged	$\mathbb{E}[\Delta\mu]$	0	σ^2_{dis}/n_{dis}
conditional-i.i.d.	0	0	σ^2_{priv}/n_{priv} (on priv), σ^2_{dis}/n_{dis} (on dis)

Empirical Results

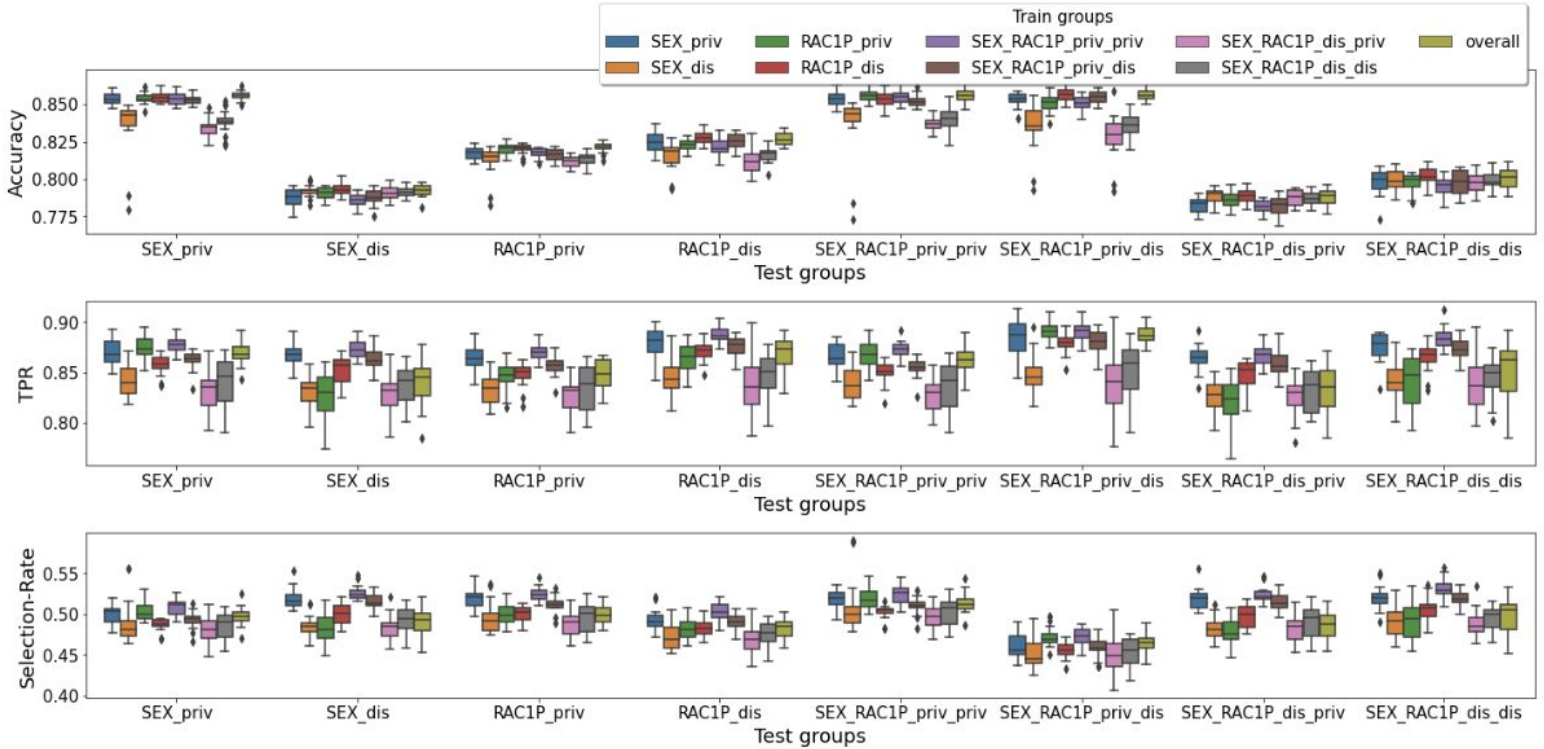


Figure 5: Conditioning on sensitive attributes, folktables: Test performance of different models broken down by test subgroup.

Next Steps

- Large-scale empirical evaluation of proposed variance metrics and conditional models
  - Accuracy-fairness-stability tradeoff
  - Benefits of using conditional models

- Good ways to combine group-specific models?

- Blind conditional models —- without conditioning on sensitive attributes