

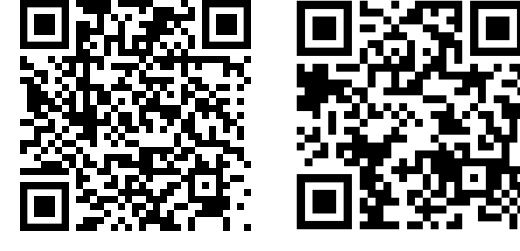
FairSHAP: Preprocessing for Fairness Through Attribution-Based Data Augmentation

Lin Zhu* Technical University of Denmark, Denmark

Yijun Bian* University of Copenhagen, Denmark

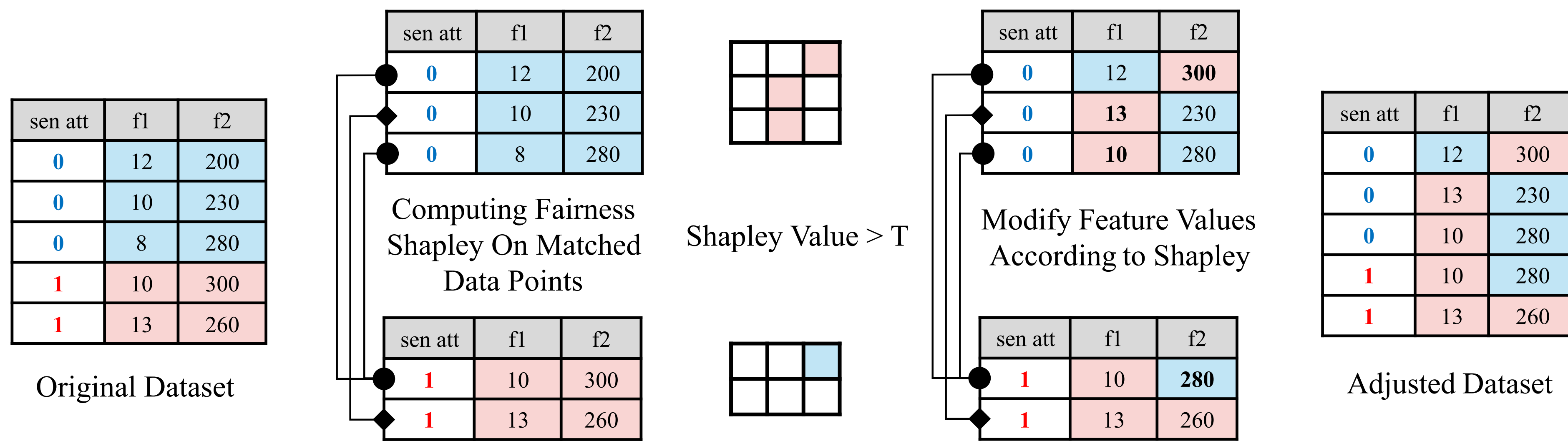
Lei You Technical University of Denmark, Denmark ✉leiyo@dtu.dk

Lin is looking for PhD positions, and Yijun is on the job market:



UNIVERSITY OF COPENHAGEN

Framework



- **Left:** Partition the original dataset by the sensitive attribute into *privileged* and *unprivileged* groups, then perform nearest-neighbour matching between two groups.
- **Middle:** Computed *Shapley value matrix* on the matched samples.
- **Right:** *Modify feature values* according to the Shapley attributions and synthesize the *adjusted dataset*.

Preliminaries and our FairSHAP

Shapley Value For players $\mathcal{F} = \{1, 2, \dots, n\}$ and $v: 2^{\mathcal{F}} \rightarrow \mathbb{R}$, the Shapley value of player k is

$$\phi_k(v) = \sum_{S \subseteq \mathcal{F} \setminus \{k\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{k\}) - v(S)). \quad (1)$$

Discriminative Risk With $\mathbf{x} = (\tilde{\mathbf{x}}, A)$, where $\tilde{\mathbf{x}} \in \mathcal{X}_{\mathcal{F} \setminus \{A\}}$ are the non-sensitive features and $A \in \{0, 1\}$ is the sensitive attribute (0 = unprivileged, 1 = privileged), DR is defined as:

$$L_{\text{DR}}(f, \mathbf{x}) = |f(\tilde{\mathbf{x}}, A = 0) - f(\tilde{\mathbf{x}}, A = 1)| \leq \epsilon. \quad (2)$$

FairSHAP For instance i , define the coalition value on features $S \subseteq \mathcal{F}$ via an individual-fairness loss, where $\mathcal{P}$ is the joint probability obtained by nearest neighbour matching:

$$\nu^{(i)}(S) = \mathbb{E}_{\tilde{\mathbf{g}} \sim \mathcal{P}(\tilde{\mathbf{g}}|\mathbf{g}_i)} [L_{\text{DR}}(\mathbf{g}_i, S; \tilde{\mathbf{g}}_{\mathcal{F} \setminus S})] - \mathbb{E}_{\tilde{\mathbf{g}} \sim \mathcal{P}(\tilde{\mathbf{g}}|\mathbf{g}_i)} [L_{\text{DR}}(\tilde{\mathbf{g}}_{\mathcal{F}})], \quad (3a)$$

$$\text{s.t. } \mathcal{P}(\mathbf{g}, \tilde{\mathbf{g}}) = \mathcal{M}_{\text{method}}(\mathcal{G}, \tilde{\mathcal{G}}). \quad (3b)$$

Algorithm

Algorithm 1 Overall framework: Enhancing fairness via matching and Shapley values

Input: Model f , dataset $\mathcal{D} \in \mathbb{R}^{(n+m) \times d}$ with sensitive attribute $A \in \{0, 1\}$, threshold T , matching method $\mathcal{M}_{\text{method}}$, where method $\in \{\text{NearestNeighbour}, \text{OptimalTransport}\}$

Output: $\mathcal{D}_{\text{new}}$

- 1: Split $\mathcal{D}$ into two subgroups: $\mathcal{G} \in \mathbb{R}^{n \times d}$ (e.g., $A = 0$) and $\tilde{\mathcal{G}} \in \mathbb{R}^{m \times d}$ (e.g., $A = 1$)
- 2: $\mathcal{G}' \leftarrow \text{FairSHAP}(\text{target} = \mathcal{G}, \text{non-target} = \tilde{\mathcal{G}}, \text{model} = f, T, \mathcal{M}_{\text{method}})$ // see Algorithm 2
- 3: $\tilde{\mathcal{G}}' \leftarrow \text{FairSHAP}(\text{target} = \tilde{\mathcal{G}}, \text{non-target} = \mathcal{G}, \text{model} = f, T, \mathcal{M}_{\text{method}})$ // see Algorithm 2
- 4: $\mathcal{D}_{\text{new}} \leftarrow \text{Concat}(\mathcal{G}', \tilde{\mathcal{G}}') \in \mathbb{R}^{(n+m) \times d}$
- 5: **return** $\mathcal{D}_{\text{new}}$

Algorithm 2 FairSHAP

Input: Target group $\mathcal{G} \in \mathbb{R}^{n \times d}$ and non-target group $\tilde{\mathcal{G}} \in \mathbb{R}^{m \times d}$, model f , threshold T , matching method $\mathcal{M}_{\text{method}}$

Output: modified dataset $\mathcal{G}'$

- 1: Use $\mathcal{M}_{\text{method}}(\mathcal{G}, \tilde{\mathcal{G}})$ to obtain joint probability $\mathcal{P}(\mathbf{g}, \tilde{\mathbf{g}}) \in \mathbb{R}^{n \times m}$
- 2: Use Eqs. (1) and (3) to obtain Shapley value matrix $\phi \in \mathbb{R}^{n \times d}$
- 3: Initialize reference data $\mathcal{B} \leftarrow \mathbf{0}_{n \times d}$
- 4: **for** $i = 1$ **to** n **do**
- 5: $j^* \leftarrow \arg\max_{1 \leq j \leq m} \mathcal{P}_{i,j}$
- 6: $\mathcal{B}_{i,:} \leftarrow \tilde{\mathcal{G}}_{j^*,:}$
- 7: Let $\mathcal{G}' \leftarrow \mathcal{G} (\mathcal{G}' \in \mathbb{R}^{n \times m})$
- 8: **for** $i = 1$ **to** n **do**
- 9: **for** $k = 1$ **to** d **do**
- 10: **if** $\phi_{i,k} \geq T$ **then**
- 11: $\mathcal{G}'_{i,k} \leftarrow \mathcal{B}_{i,k}$
- 12: **return** $\mathcal{G}'$

Qualitative Results

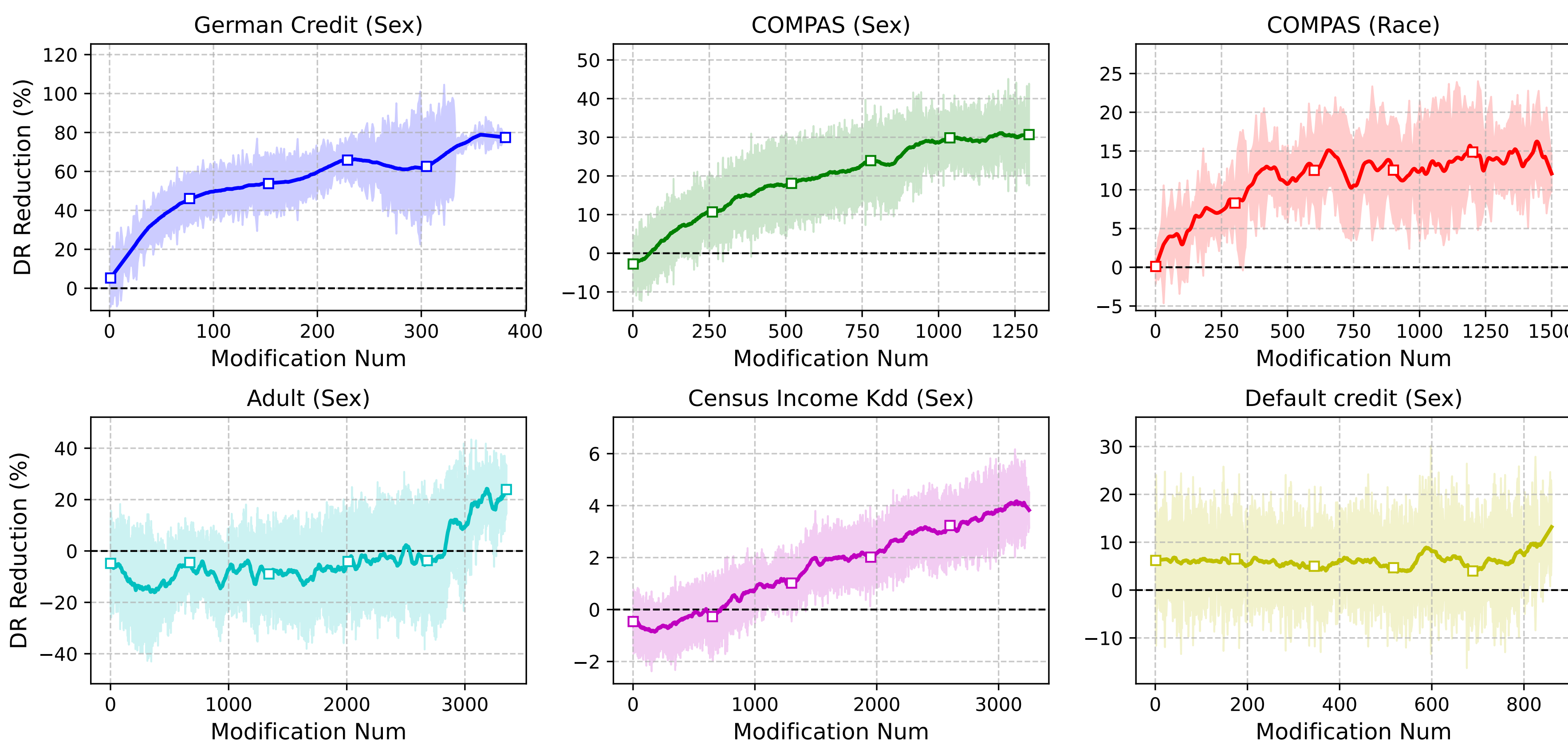


Figure 2: Percentage reduction in the discriminative risk (DR) across different datasets.

Quantitative Results

Table 1: Compare FairSHAP with other fairness mitigation methods across different datasets. (CR: CorrelationRemover; DIR: DisparateImpactRemover)

Dataset (s.a.)	Methods	Accuracy	DR	DP	EO	PQP	Data Fidelity	TrainingAR	TestAN
German (sex)	Baseline	0.6650±0.0257	0.0785±0.0211	0.0512±0.0346	0.1287±0.0590	0.1341±0.0486	—	—	No
	CR	0.6680±0.0238	0.0028±0.0029	0.0844±0.0557	0.1559±0.0609	0.0723±0.0330	0.0183±0.0211	0.9615	Yes
	DIR	0.6720±0.0337	0.0966±0.0112	0.0946±0.0373	0.1737±0.0729	0.1529±0.0634	0.0155±0.0440	0.0774	Yes
	FairSHAP	0.6630±0.0275	<u>0.0243±0.0112</u>	0.0301±0.0347	0.1126±0.0783	0.1852±0.1074	0.0049±0.0085	0.0156	No
COMPAS (sex)	Baseline	0.6698±0.0051	0.0883±0.0064	0.1548±0.0241	0.1243±0.0510	0.0492±0.0084	—	—	No
	CR	0.6679±0.0045	0.0082±0.0070	0.1407±0.0248	0.1291±0.0317	0.0714±0.0517	0.0189±0.0193	0.9174	Yes
	DIR	0.6644±0.0098	0.1150±0.0091	0.1155±0.0239	0.0952±0.0359	0.0747±0.0370	0.0387±0.0640	0.0650	Yes
	FairSHAP	0.6609±0.0106	<u>0.0629±0.0091</u>	<u>0.1326±0.0407</u>	<u>0.0985±0.0603</u>	0.0452±0.0383	0.0025±0.0048	0.0113	No
COMPAS (race)	Baseline	0.6689±0.0108	0.0995±0.0076	0.1436±0.0209	0.1438±0.0233	0.0522±0.0406	—	—	No
	CR	0.6611±0.0112	0.0418±0.0092	0.1502±0.0341	0.1621±0.0530	0.0592±0.0367	0.0250±0.0222	0.892	Yes
	DIR	0.6149±0.0286	0.1185±0.0181	<u>0.1359±0.1241</u>	0.1117±0.0945	0.0399±0.0338	0.0512±0.0736	0.0701	Yes
	FairSHAP	0.6627±0.0069	<u>0.0842±0.0049</u>	0.1344±0.0332	0.1568±0.0343	<u>0.0508±0.0469</u>	0.0040±0.0055	0.0126	No
Adult (sex)	Baseline	0.8722±0.0033	0.0315±0.0037	0.1805±0.0066	<u>0.0735±0.0275</u>	0.0275±0.0321	—	—	No
	CR	0.8706±0.0029	0.0000±0.0000	0.1824±0.0055	0.0955±0.0243	0.0278±0.0173	0.0167±0.0391	0.9887	Yes
	DIR	0.8550±0.0067	0.0499±0.0076	<u>0.1607±0.0157</u>	0.0772±0.0624	0.0360±0.0253	0.0046±0.0417	0.0081	Yes
	FairSHAP	0.8692±0.0046	<u>0.0273±0.0047</u>	0.1558±0.0130	0.0393±0.0254	0.0474±0.0319	0.0010±0.0073	0.0012	No
Adult (race)	Baseline	0.8721±0.0033	0.0398±0.0025	0.1034±0.0110	<u>0.0808±0.0326</u>	<u>0.0302±0.0265</u>	—	—	No
	CR	0.8713±0.0033	0.0000±0.0000	0.1008±0.0115	0.0983±0.0389	0.0480±0.0235	0.0300±0.0450	0.962	Yes
	DIR	0.8320±0.0173	0.0740±0.0209	0.0703±0.0515	0.0871±0.0355	0.0482±0.0730	0.0252±0.0480	0.0089	Yes
	FairSHAP	<u>0.8720±0.0023</u>	<u>0.0284±0.0017</u>	<u>0.0851±0.0155</u>	0.0287±0.0277	0.0259±0.0318	0.0030±0.0084	0.0014	No