Exploiting Fairness to Enhance Sensitive Attributes Reconstruction

Julien Ferry[1], Ulrich Aïvodji[2], Sébastien Gambs[3], Marie-José Huguet[1] and Mohamed Siala[1]

[1]LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France
[2]École de Technologie Supérieure, Montréal, Canada
[3]Université du Québec à Montréal, Montréal, Canada

February 23, 2023

Table of Contents

Classification

- Consider some high-stakes decision making task, such as *college admissions*
- Consider a labeled dataset $D = (X, S, Y) \in (\mathcal{X} \times \mathcal{S} \times \mathcal{Y})^N$ such that:
  - X is a set of unsensitive attributes, which can be used for decision making (*e.g., high school grades*)
  - S is a set of sensitive attributes, which should not be used for decision making (*e.g., gender*)
  - Y is the ground truth label (*e.g., admission decision (yes/no)*)

Statistical Fairness Metrics: Principle

- Several *protected subgroups* defined by the different values of the sensitive attributes
- *Statistical/Group Fairness:* Ensure that some statistical measure $\mathcal{M}$ of a classifier's h outputs differs by no more than a given *tolerance* ϵ between the different *protected groups* and the overall dataset

Fair Learning Problem

- A fair learning procedure $\mathcal{L}$ aims to produce a fair classifier $h : \mathcal{X} \mapsto \mathcal{Y}$ minimizing some objective function obj(.) over some hypothesis space $\mathcal{H}$

Statistical measure (*e.g.*, positive prediction rate for the Statistical Parity fairness metric)

$$\underset{h \in \mathcal{H}}{\arg\min} \quad \mathrm{obj}(h, D)$$
$$\text{s.t. } \forall s \in \mathcal{S}, \quad |\boxed{\mathcal{M}}(h, \{e \in D\}) - \boxed{\mathcal{M}}(h, \{e \in D \mid s_e = s\})| \leq \boxed{\epsilon}$$

Unfairness tolerance

Reconstruction Attacks

- *Reconstruction Attacks* are a type of *inference attack* first proposed against database access mechanisms [Dinur and Nissim, 2003]
  - An adversary knows an entire database except one private column, and tries to reconstruct it

Sensitive Attributes Reconstruction Attack

- An adversary with some auxiliary knowledge (*e.g.*, (X, Y)) has black-box access to fair model h trained on D
- The adversary wants to reconstruct the training set sensitive attributes column S (which h does not use for decision-making)

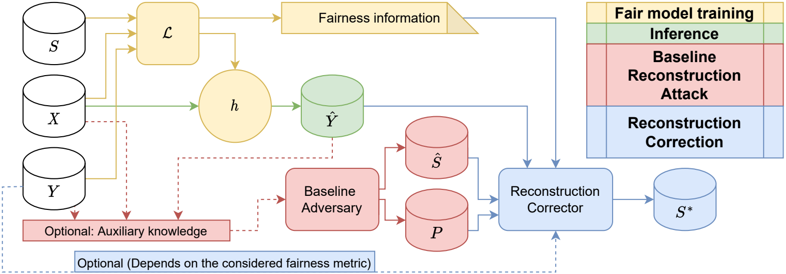

Figure: The proposed attack framework.

1. A model h is learnt by the fair learning procedure $\mathcal{L}$ and used for inference
2. A *Baseline Adversary* tries to reconstruct the sensitive attributes S of h's training set
3. Our proposed *Reconstruction Corrector* component takes as input the Baseline Adversary's reconstruction $\hat{S}$ and corrects it to comply with the fairness information

The Integer Programming Model $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$

- Inputs:
  - $\hat{s}_i \in \{0,1\}, \ i = 1, \ldots, N$ (baseline adversary's reconstruction)
  - $p_i \in \{0,1\}, \ i = 1, \ldots, N$ (adversary's confidence for $\hat{s}_i$)
  - $\hat{y}_i \in \{0,1\}, \ i = 1, \ldots, N$ (target model h's predictions)
  - Fairness information: h satisfies fairness constraints for some metric (*e.g.*, SP) and some tolerance ϵ
- Decision variables:
  - $s_i^* \in \{0,1\}, \ i = 1, \ldots, N$ (corrected sensitive attributes reconstruction)

Confidence-weighted #changes to $\hat{S}$ →

$$\min \sum_{i=1}^{N} (p_i \cdot (1 - \hat{s}_i) \cdot s_i^*) + \sum_{i=1}^{N} (p_i \cdot \hat{s}_i \cdot (1 - s_i^*)) \tag{1}$$

At least one example in each group →

$$s.t.: \quad 0 < \sum_{i=1}^{N} s_i^* < N \tag{2}$$

Group 1 fairness constraint →

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{\sum_{i=1}^{N} \hat{y}_i \cdot s_i^*}{\sum_{i=1}^{N} s_i^*} \leq \epsilon \tag{3}$$

Group 0 fairness constraint →

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{\sum_{i=1}^{N} \hat{y}_i \cdot (1 - s_i^*)}{\sum_{i=1}^{N} (1 - s_i^*)} \leq \epsilon \tag{4}$$

Pros and Cons

- (+) Can encode any constraint over the sensitive attributes
- (-) Exponential search space (w.r.t. the number of examples N): not scalable
  - For statistical fairness constraints: don't need such granularity as only counts (per protected group/per prediction/per label) matter

The Constraint Programming Model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$

- Inputs:
  - Baseline reconstruction cardinalities $n_{\mathbf{1}}^{+}$, $n_{\mathbf{0}}^{+}$, $n_{\mathbf{1}}^{-}$ and $n_{\mathbf{0}}^{-}$
  - Arrays of sorted and cumulated adversary's confidences for each example's baseline reconstruction: $T_{\mathbf{1}^{+}}$, $T_{\mathbf{0}^{+}}$, $T_{\mathbf{1}^{-}}$ and $T_{\mathbf{0}^{-}}$
  - Fairness information: h satisfies fairness constraints for some metric (*e.g.*, SP) and some tolerance ϵ
- Decision variables:
  - $s_{\mathbf{01}}^{+} \in [0, n_{\mathbf{0}}^{+}]$: number of changes of $\hat{s}_i$ from 0 to 1, for examples such that $\hat{y}_i = 1$
  - $s_{\mathbf{10}}^{+} \in [0, n_{\mathbf{1}}^{+}]$: number of changes of $\hat{s}_i$ from 1 to 0, for examples such that $\hat{y}_i = 1$
  - $s_{\mathbf{01}}^{-} \in [0, n_{\mathbf{0}}^{-}]$: number of changes of $\hat{s}_i$ from 0 to 1, for examples such that $\hat{y}_i = 0$
  - $s_{\mathbf{10}}^{-} \in [0, n_{\mathbf{1}}^{-}]$: number of changes of $\hat{s}_i$ from 1 to 0, for examples such that $\hat{y}_i = 0$
- For instance, consider that re-establishing fairness requires to swap five positively predicted examples' sensitive attributes from 0 to 1
  - Then, $s_{01}^{+} = 5$ and $T_{0^{+}}[s_{01}^{+}]$ is the cost of changing the sensitive attribute value from 0 to 1 for five examples positively predicted

The Constraint Programming Model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$

Confidence-weighted #changes to $\hat{S}$ $\rightarrow$ min $T_{0^+}[s^+_{01}] + T_{1^+}[s^+_{10}] + T_{0^-}[s^-_{01}] + T_{1^-}[s^-_{10}]$ (5)

At least one example in group 0 $\rightarrow$ s.t. : $n^+_0 + n^-_0 - s^+_{01} - s^-_{01} + s^+_{10} + s^-_{10} > 0$ (6)

At least one example in group 1 $\rightarrow$ $n^+_1 + n^-_1 - s^+_{10} - s^-_{10} + s^+_{01} + s^-_{01} > 0$ (7)

Group 1 fairness constraint $\rightarrow$ $-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{n^+_1 - s^+_{10} + s^+_{01}}{n^+_1 + n^-_1 - s^+_{10} - s^-_{10} + s^+_{01} + s^-_{01}} \leq \epsilon$ (8)

Group 0 fairness constraint $\rightarrow$ $-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{n^+_0 - s^+_{01} + s^+_{10}}{n^+_0 + n^-_0 - s^+_{01} - s^-_{01} + s^+_{10} + s^-_{10}} \leq \epsilon$ (9)

Pros and Cons

- (+) Polynomial search space (w.r.t. the number of examples N): scalable
  - Solved to optimality in fractions of seconds in all our experiments with $N > 100,000$
- (-) Can only encode group-level constraints over the sensitive attributes

Setup Description: Learning Fair (Target) Models

- (Target) Fair models are learnt using the `Fairlearn` library [Bird et al., 2020]
- A wide range of unfairness tolerances with four fairness metrics:
  - **Statistical Parity** [Dwork et al., 2012]
  - Predictive Equality [Chouldechova, 2017]
  - Equal Opportunity [Hardt et al., 2016]
  - Equalized Odds [Hardt et al., 2016]
- Three biased datasets with diverse characteristics:

Dataset	Binary Prediction Task	#Datapoints	#Non-Sensitive Features	Sensitive Feature
UCI Adult Income [Dua and Graff, 2017]	Income above $50K	45,222	7 categorical, 6 numerical	Gender (Male/Female)
ACSPublicCoverage* [Ding et al., 2021]	Coverage from public health insurance	98,928	17 categorical, 1 numerical	Age (First Quartile/Others)
ACSIncome* [Ding et al., 2021]	Income above $50K	135,924	7 categorical, 2 numerical	Race Code (White/Other)

*(Texas State, 2018)

Setup Description: Reconstruction Attack

- Baseline Adversary's Reconstruction: ML-based adversary proposed by Aalmoes et al. [2022], as informed as our Reconstruction Corrector component
  - $\Longrightarrow$ **Strongest baseline possible**
- Corrected Reconstruction: our proposed efficient model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ is solved using `IBM ILOG CP Optimizer` and its default configuration

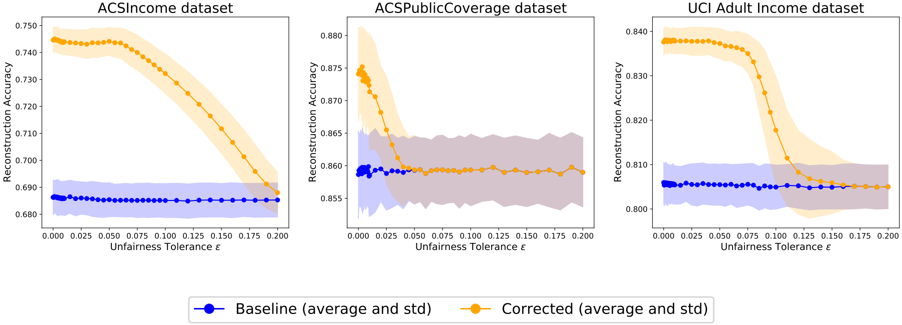

Figure: Baseline and corrected reconstruction quality, for our experiments using the Statistical Parity metric

Additional Contributions

- The attack success does not depend on the type of fairness intervention (pre-processing, in-processing, post-processing) as black-box access to the model's predictions are sufficient
- Even if it is not revealed explicitly, the fairness information can be inferred and the attack still succeeds (and sometimes, even perform better!)
- Considering a weaker baseline adversary, baseline reconstruction performances are lower but our reconstruction correction step provides accuracy improvements of comparable magnitude

Summary

- We propose a novel approach to improve sensitive attributes reconstruction by a baseline adversary by incorporating user-defined constraints
- We introduce two models implementing such approach, with genericity or efficiency advantages
- Our results show that the fairness information can be leveraged to improve the success of sensitive attributes reconstruction attacks

Future Work

- Combining our attack with different baseline adversaries
- Applying our framework in the context of multi-valued sensitive attributes

Contact

- Email: `jferry@laas.fr`
- Homepage: `https://homepages.laas.fr/jferry/`

Links

- Full paper @SATML 2023 - *The 1st IEEE Conference on Secure and Trustworthy Machine Learning*: `https://openreview.net/forum?id=t0Vr0HLaFz0`
- Source code: `https://github.com/ferryjul/SensitiveAttributesReconstructionCorrector/`

References I

Aalmoes, J., Duddu, V., and Boutet, A. (2022). Dikaios: Privacy auditing of algorithmic fairness via attribute inference attacks. *arXiv preprint arXiv:2202.02242*.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. (2018). A reductions approach to fair classification. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

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

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 6478–6490.

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In Neven, F., Beeri, C., and Milo, T., editors, *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210. ACM.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

- We know that h is fair on its training dataset $D = (X, S, Y)$
- Yet, the reconstruction $\hat{S}$ outputted by some baseline adversary may not comply with the fairness information
- Then, if h is not fair on $(X, \hat{S}, Y)$, we know that $\hat{S} \neq S$
- $\implies$ We post-process $\hat{S}$ to compute S^*, a corrected version complying with the fairness information and minimizing the confidence-weighted changes to $\hat{S}$

Setup Description: Reconstruction Attack

- (Target) Fair models are learnt using the `ExponentiatedGradient` [Agarwal et al., 2018] fair in-processing method (`Fairlearn` library [Bird et al., 2020]) with `scikit-learn` [Pedregosa et al., 2011] `DecisionTreeClassifiers` as base learners
- Baseline Adversary Original Reconstruction: ML-based adversary proposed in [Aalmoes et al., 2022], as informed as our Reconstruction Corrector component
  - Adversarial Knowledge:
    - Auxiliary attack set $D_A = (X_A, S_A, Y_A)$
    - Training set non-sensitive attributes vector and true labels (X, Y)
    - Black-box access to the target fair model h
  - Description of the Attack:
    1. Computes $\hat{Y}_A = h(X_A)$ and $\hat{Y} = h(X)$
    2. Trains a machine learning model (coined *attack model*) to predict S_A from $(X_A, Y_A, \hat{Y}_A)$ (we tune the attack model's hyperparameters using a validation set)
    3. Uses its trained *attack model* to predict $(\hat{S}, P)$ from $(X, Y, \hat{Y})$
- Corrected Reconstruction: our proposed efficient model $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ is solved using the `IBM ILOG CP Optimizer` via the `DOcplex` Python Modeling API and its default configuration. It outputs a corrected reconstruction S^*

Results using an In-Processing Method for Fairness I

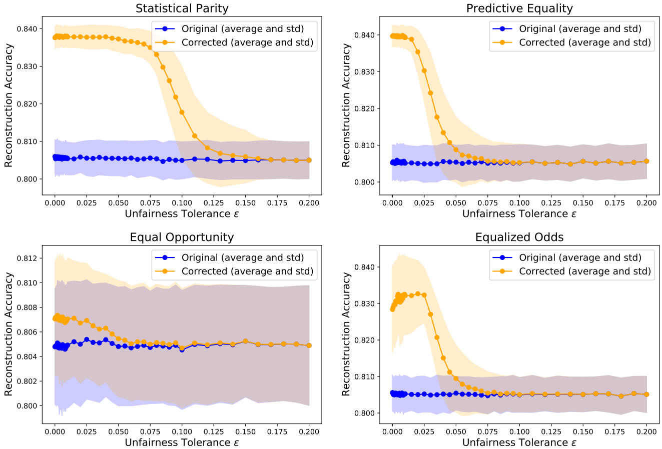

Figure: Corrected and original reconstruction quality, for our experiments using the UCI Adult Income dataset.

Results using an In-Processing Method for Fairness II

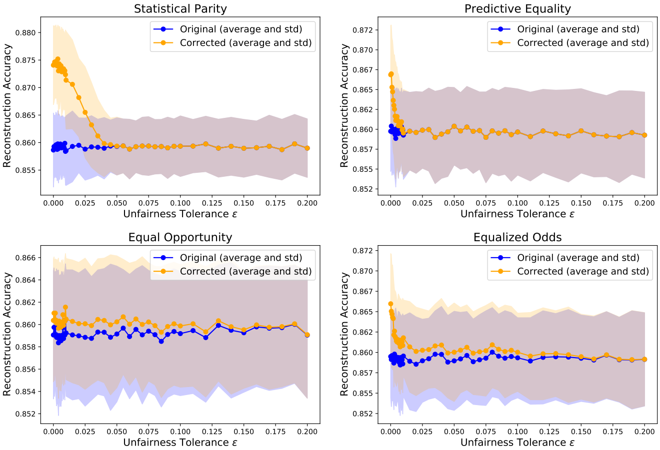

Figure: Corrected and original reconstruction quality, for our experiments using the ACSPublicCoverage dataset.

Results using an In-Processing Method for Fairness III

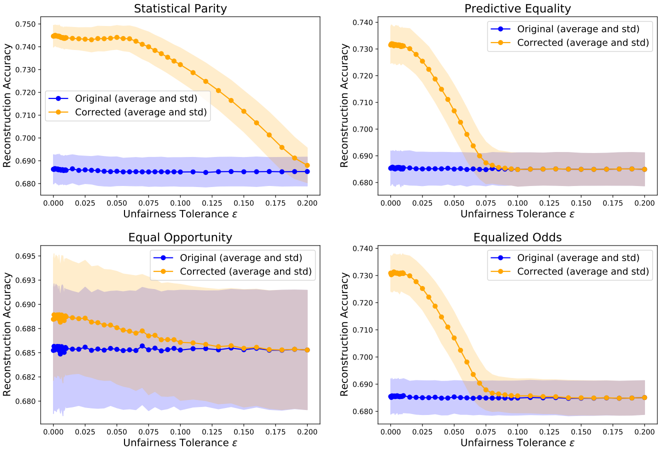

Figure: Corrected and original reconstruction quality, for our experiments using the ACSIncome dataset

Results using an In-Processing Method for Fairness (Countermeasure) I

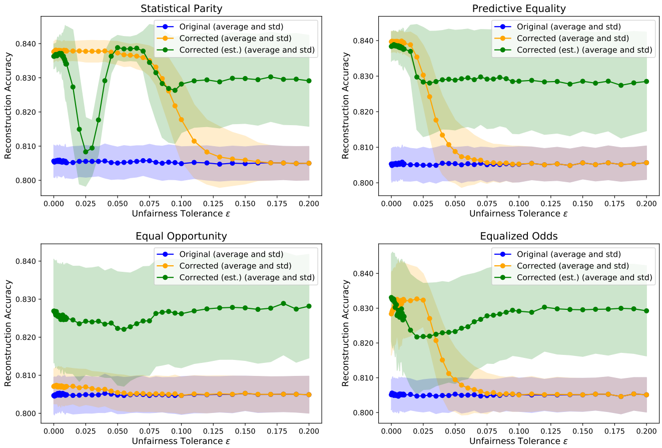

Figure: Original, corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the UCI Adult Income dataset

Results using an In-Processing Method for Fairness (Countermeasure) II

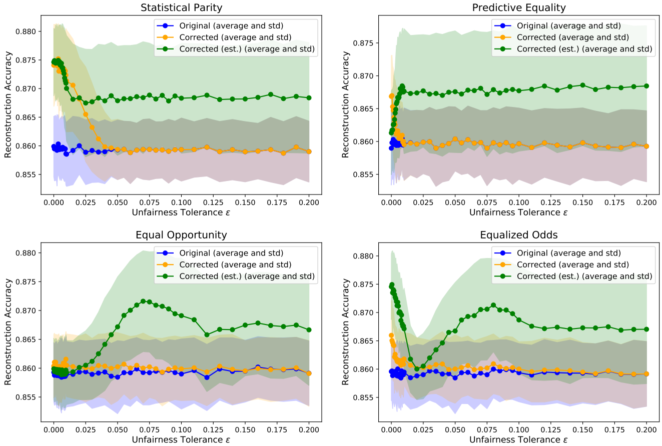

Figure: Original, corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the ACSPublicCoverage dataset

Results using an In-Processing Method for Fairness (Countermeasure) III

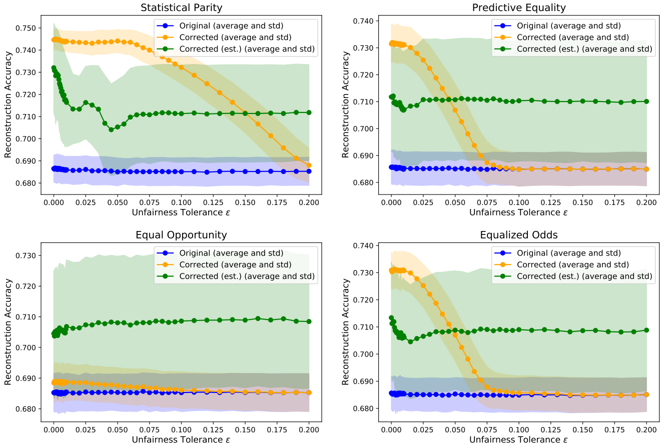

Figure: Original, corrected (from actual fairness constraint, and from estimated one (`est.`)) reconstruction quality, for our experiments using the ACSIncome dataset

Results using a Post-Processing Method for Fairness

Table: Summary of the results of our experiments using the `ThresholdOptimizer` [Hardt et al., 2016] fair post-processing method implemented in the `Fairlearn` library [Bird et al., 2020]

Metric	Reconstruction Perf. *Original*	Reconstruction Perf. *Corrected*	Estimated Constraint *Metric Detect.*	Estimated Constraint *Average Tolerance*	Reconstruction Perf. (Corrected from Estimated Constraint)
UCI Adult Income dataset					
SP	0.814 ± 0.006	0.858 ± 0.005	0.95	0.004 ± 0.003	0.856 ± 0.011
PE	0.807 ± 0.005	0.844 ± 0.004	0.97	0.003 ± 0.002	0.843 ± 0.007
EO	0.805 ± 0.005	0.807 ± 0.005	0.26	0.018 ± 0.010	0.828 ± 0.013
EOdds	0.807 ± 0.004	0.840 ± 0.009	0.00	0.005 ± 0.005	0.843 ± 0.007
ACSPublicCoverage dataset					
SP	0.860 ± 0.006	0.875 ± 0.007	1.00	0.002 ± 0.002	0.873 ± 0.009
PE	0.860 ± 0.005	0.870 ± 0.007	1.00	0.003 ± 0.002	0.865 ± 0.007
EO	0.859 ± 0.006	0.861 ± 0.006	0.28	0.008 ± 0.005	0.862 ± 0.005
EOdds	0.860 ± 0.005	0.861 ± 0.005	0.00	0.002 ± 0.002	0.869 ± 0.007
ACSIncome dataset					
SP	0.715 ± 0.010	0.764 ± 0.006	0.80	0.003 ± 0.003	0.754 ± 0.020
PE	0.688 ± 0.007	0.735 ± 0.006	0.86	0.003 ± 0.003	0.728 ± 0.016
EO	0.685 ± 0.006	0.689 ± 0.006	0.73	0.008 ± 0.006	0.700 ± 0.020
EOdds	0.688 ± 0.007	0.735 ± 0.006	0.00	0.002 ± 0.002	0.721 ± 0.022

Results using a Pre-Processing Method for Fairness

Table: Summary of the results of our experiments using the `CorrelationRemover` fair pre-processing method implemented in the `Fairlearn` library [Bird et al., 2020]

Target model (under attack)		Estimated Constraint		Reconstruction Perf.	
Train Acc.	*Test Acc.*	*Estimated Metric*	*Estimated Tolerance*	*Original*	*Corrected*
UCI Adult Income dataset					
0.860 ± 0.003	0.848 ± 0.003	PE (68%), EO (32%)	0.023 ± 0.013	0.806 ± 0.005	0.827 ± 0.014
ACSPublicCoverage dataset					
0.862 ± 0.001	0.852 ± 0.002	PE (92%), SP (8%)	0.006 ± 0.004	0.860 ± 0.006	0.872 ± 0.010
ACSIncome dataset					
0.798 ± 0.002	0.785 ± 0.003	PE (100%)	0.056 ± 0.016	0.685 ± 0.008	0.763 ± 0.009