Counterfactual Reasoning for Decision Model Fairness Assessment

Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Eugenio Di Sciascio
firstname.lastname@poliba.it
Polytechnic University of Bari
Bari, Italy

Azzurra Ragone
azzurra.ragone@uniba.it
University of Bari
Bari, Italy

ABSTRACT

The increasing application of Artificial Intelligence and Machine Learning models poses potential risks of unfair behaviour and, in the light of recent regulations, has attracted the attention of the research community. Several researchers focused on seeking new fairness definitions or developing approaches to identify biased predictions. These approaches focus solely on a discrete and limited space; only a few analyze the minimum variations required in the user characteristics to ensure a positive outcome for the individuals (counterfactuals). In that direction, the methodology proposed in this paper aims to unveil unfair model behaviors using counterfactual reasoning in the case of fairness under unawareness. The method also proposes two new metrics that analyse the (estimated) sensitive information of counterfactual samples with the help of an external oracle. Experimental results on three data sets show the effectiveness of our approach for disclosing unfair behaviour of state-of-the-art Machine Learning and debiasing models. Source code is available at https://github.com/giandos200/WWW-23-Counterfactual-Fair-Opportunity-Poster-.

CCS CONCEPTS

• **Social and professional topics** → **Gender**; **Race and ethnicity**.

KEYWORDS

Bias, Fairness, Audit, Counterfactual Reasoning

1 INTRODUCTION

Artificial Intelligence (AI) systems are increasingly pervasive in our society and often exploited for taking life-changing decisions, like loans, job offers, and health care access. One of the inherent risks linked to those tasks is the *discrimination* of groups or individuals.

In the fintech industry, *online instant lending platforms* use machine learning tools to analyze available consumer credit data to make faster credit decisions. Nevertheless, in the financial sector, the choice to grant or deny a credit has been regulated by rigorous and thorough regulatory compliance criteria referring primarily to human-decision (e.g., Equal Credit Opportunity Act and Consumer Credit Directive for EU Community). However, when AI replaces human decisions, like in the case of instant lending, there is a risk of revealing a loophole in existing liability identification laws. Several national and international organizations have released guidelines, norms, and principles to prevent the irresponsible usage of AI, e.g., the EU Commission with "The Proposal for Harmonized Rule on AI" and the expert group on "AI in Society" of the Organisation for Economic Co-operation and Development (OECD).

Although scientists train their models without explicit discriminating intent, deploying AI systems without taking ethical concerns into account may lead to discrimination [2]. Even more problematic is figuring out which type of discrimination is being implemented. In the last years, a wide range of definitions has been proposed for fairness [11]. The scientific community has drawn up a wide range of fairness definitions that are derived from specific legal, philosophical, or mathematical applications. Unfortunately, since the most often used criteria for fairness frequently conflict with one another, if we make an algorithm fair on one measure, it could become unfair on another [6, 9]. Generally, fairness definitions refer to people separable into privileged and unprivileged group, characterized by sensitive information (e.g. gender, age, race) within which metrics of disparity in outcome are measured, i.e., *group-fairness*. Dwork et al. [5] introduced the concept of *individual-fairness* according to which similar individuals should be treated similarly. Between the two definitions and particularly in the latter lies the basis for the *Counterfactual Fairness* [7]: '*a predictor can be considered counterfactually fair if its result does not change between individual with the same characteristics but different sensitive information*'. This definition requires the use of sensitive features that in particular domains (e.g. finance, health care) is usually forbidden. Moreover, removing sensitive information does not ensure the predictor fair behavior [3], and it prevents an ex-post auditing of the model fairness.

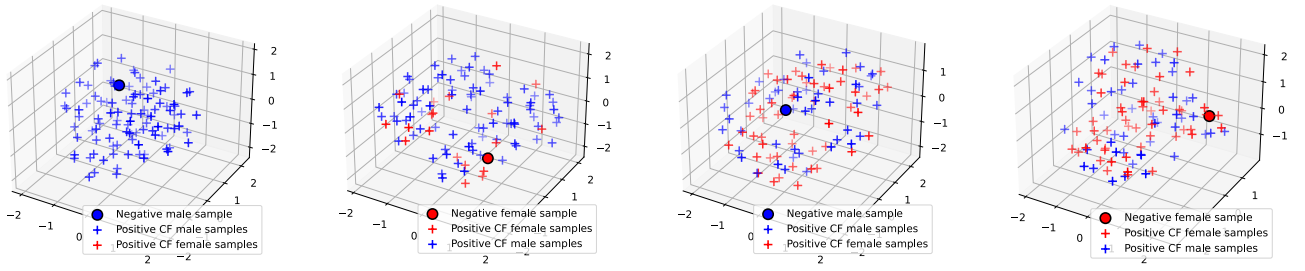
Our work overcomes this limitation being able to detect bias in the case of *fairness under unawareness* [1] through the use of counterfactual reasoning [10], exploiting two novel fairness metrics. The proposed metrics, i.e., *Counterfactual Flips* (CFlips) and *normilized Discounted Cumulative Counterfactual Fairness* (nDCCF) identify the discriminatory behavior of the **Decision Maker** using a **Countrerfactual Generator** and an oracle (i.e., **Sensitive-Feature Classifier**). The metrics explore a new fairness criteria: *Counterfactual Fair Opportunity*. Our auditing methodology aims to be an effective tool for quantifying the discriminatory behavior of any ML model.

2 PRELIMINARIES

This section introduces the notation adopted hereinafter.

WWW '23, April 30–May 04, 2023, Austin, Texas, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX



(a) male on Classic ML model **(b) female on Classic ML model** **(c) male on Debiasing model** **(d) female on Debiasing model**

Figure 1: Adult t-SNE visualizations of a random male (a-c) and female (b-d) sample with a negative outcome and their CF samples with a positive outcome, respectively, for a Classic ML model (i.e. XGB) and a Debiasing model (i.e. Adversarial Debiasing).

Data points: We assume the dataset $\mathcal{D}$ is an m-dimensional space containing n non-sensitive features, l sensitive features, and a target attribute. In other words, we have $\mathcal{D} \subseteq \mathbb{R}^m$, with $m = n + l + 1$[1]. A data point $d \in \mathcal{D}$ is then represented as $d = \langle \mathbf{x}, \mathbf{s}, y \rangle$, with $\mathbf{x} = \langle x_1, x_2, ..., x_n \rangle$ representing the sub-vector of non-sensitive features, $\mathbf{s} = \langle s_1, s_2, ..., s_l \rangle$ the sub-vector of sensitive features and y being a binary target feature. Given a vector of sensitive festures, $\forall s_i \in \mathbf{s}$, $s_i = 0$ refers to the *unprivileged* group and $s_i = 1$ to the *privileged* group of the i-th sensitive feature.
Target Labels: Given a target feature $y \in \{0, 1\}$, $y = 1$ is the positive outcome and $y = 0$ is the negative one.
Outcome Prediction: $\hat{y} \in \{0, 1\}$ represents the prediction for $\mathbf{x} \subset d$ estimated by $f(\cdot)$, a function such that $f(\mathbf{x}) = \hat{y}$.
Sensitive Feature Prediction: $\hat{s}_i \in \{0, 1\}$ represents the prediction of the i-th sensitive feature for a given data point estimated by $f_{s_i}(\cdot)$, a function s.t. $f_{s_i}(\mathbf{x}) = \hat{s}_i$.
Counterfactual samples: Given a vector $\mathbf{x}$ and a perturbation $\epsilon = \langle \epsilon_1, \epsilon_2, ..., \epsilon_n \rangle$, we say that a vector $\mathbf{c_x} = \langle c_{x_1}, c_{x_2}, ..., c_{x_n} \rangle = \mathbf{x} + \epsilon$ is a counterfactual (CF) of $\mathbf{x}$ if $f(\mathbf{c_x}) = 1 - f(\mathbf{x}) = 1 - \hat{y}$. We use the set $C_\mathbf{x}$, with $|C_\mathbf{x}| = k$, to denote the set of possible **counterfactual samples** for $\mathbf{x}$. A function $g(\mathbf{x})$ compute k counterfactuals for $\mathbf{x}$.
For simplicity, we denote $f(\cdot)$, $f_{s_i}(\cdot)$, and $g(\cdot)$ as the **Decision Maker**, the **Sensitive-Feature Classifier**, and the **Counterfactual Generator** respectively.

3 METHODOLOGY

Our study proposes two novel metrics for detecting bias in a scenario where sensitive features are omitted (i.e., *fairness under unawareness*) in the training process. Excluding sensitive features makes verifying that all users are treated equally incredibly challenging. In the instant lending case, imagine that a customer applies for a loan, and his/her request is rejected. Understanding if the customer has been discriminated is hard to verify when sensitive information is not used. Our process pipeline is as follows: the **Decision Maker** makes decisions without exploiting sensitive features, then if the outcome is negative (e.g. loan rejected), the **Counterfactual Generator** is exploited to propose modifications to user characteristics and request for reaching a positive outcome (e.g. loan approved). For each data point d with a negative prediction $f(\mathbf{x}) = 0$, we generate a set of counterfactual samples $C_\mathbf{x}$ that reach a positive outcome (i.e., $\forall \mathbf{c_x} \in C_\mathbf{x}$ s.t. $f(\mathbf{c_x}) = 1$). Afterward, each counterfactual (CF) sample is evaluated by the **Sensitive-Feature Classifier** that predicts the value of the (omitted) sensitive feature for the given CF sample. If the CF sample is classified as e.g. male (privileged group), while the original sample was e.g. female (unprivileged group), the decision model could be biased and its unfairness can be quantified (Eq. 3 and 4).

Indeed, each CF sample derives from the original sample $\mathbf{x}$ plus a perturbation ϵ, where ϵ is the *distance* from the original sample for getting a positive outcome, and it should be independent from the user-sensitive characteristics. Figure 1 depicts a scenario in which *male* (blu color) is the privileged category, and *female* (red color) is the unprivileged one. For each subfigure, a sample with an unfavorable decision and its corresponding CFs are depicted. A classic ML model (i.e., XGB) is compared with a debiasing ML model (i.e., AdvDeb). We can observe that for the male sample and classic ML model (Figure 1(a)), the CF samples belong to the same sensitive category (i.e., male). For the female sample (Figure 1 (b)), this is not true, revealing a bias of the model. Conversely, the debiasing model (Figure 1 (c) and (d)) shows no predominance in the generated counterfactuals of one value of the sensitive class. However, a change of the outcome, e.g. from negative to positive, should not be determined by a flip of the value(s) of the sensitive feature(s). Now, we introduce our fairness criteria and metrics.

Definition 3.1 (Counterfactual Fair Opportunity)**.** *A decision model is fair if the counterfactual samples of individuals with unfavorable decisions maintain the same sensitive value to reach a positive outcome. This behavior must be guaranteed both for the privileged and the unprivileged group.*

$$\mathbb{P}(f_s(C_{X|^-_{s=0}}) \neq s \mid f(C_{X|^-_{s=0}}) = 1, X|^-_{s=0}) = \mathbb{P}(f_s(C_{X|^-_{s=1}}) \neq s \mid f(C_{X|^-_{s=1}}) = 1, X|^-_{s=1}) \tag{1}$$

To define a sort of discrimination score of a given decision model, we propose a metric that we call *Counterfactual Flips*. The metric quantifies the discriminatory behavior the model might put in place.

Definition 3.2 (Counterfactual Flips)**.** *Given a sample* $\mathbf{x}$ *belonging to a demographic group* s *whose model output is denoted as* $f(\mathbf{x})$*, a generated set* $C_\mathbf{x}$ *of* k *counterfactuals with desired* y^* *outcome.* $\forall \mathbf{c}^i_\mathbf{x} \in C_\mathbf{x}$ *s.t.* $f(\mathbf{c}^i_\mathbf{x}) = y^*$*, the Counterfactual Flips indicate the percentage of counterfactual samples belonging to another demographic group (i.e.,* $f_s(\mathbf{c}^i_\mathbf{x}) \neq f_s(\mathbf{x})$*, with* $f_s(\mathbf{x}) = s$*).*

$$\text{CFlips}(\mathbf{x}, C_\mathbf{x}, f_s(\cdot)) \triangleq \frac{\sum_{i=1}^{k}(\mathbb{1}(\mathbf{c}^i_\mathbf{x}))}{k} \quad \textit{where } \mathbb{1}(\mathbf{c}^i_\mathbf{x}) = \begin{cases} 1 & \textit{if } f_s(\mathbf{c}^i_\mathbf{x}) \neq f_s(\mathbf{x}) \neq s \\ 0 & \textit{if } f_s(\mathbf{c}^i_\mathbf{x}) = f_s(\mathbf{x}) = s \end{cases} \tag{2}$$

[1]Without loss of generality, we assume that categorical features can always be transformed into features in $\mathbb{R}$ via one-hot-encoding.

The bigger the CFlips value is, the stronger the bias the model suffers from. In our work, we only take into account samples negatively predicted by the decision maker (i.e., $f(\mathbf{x}) = 0$) as we are interested in quantifying the discrimination in achieving a positive counterfactual result (i.e., $f(\mathbf{c_x}) = 1 \wedge f_s(\mathbf{c_x}) \neq s$). Given a set of samples $\mathcal{X}^- \subseteq \mathcal{D}$ predicted by the decision maker as negative (unfavorable decision), the metric in Eq. 2 can be generalized to the *unprivileged* and *privileged* group (in Eq. 3 $s = 0$ for the *unprivileged* samples negatively predicted, and $s = 1$ for the *privileged* samples negatively predicted).

$$\text{CFlips}_s \triangleq \frac{\sum_{i=1}^{n} \text{CFlips}(\mathbf{x}_i, C_{\mathbf{x}_i}, f_s(\cdot))}{|\mathcal{X}|_s^-|} \quad \text{with } \mathbf{x}_i \in \mathcal{X}|_s^- \tag{3}$$

A limitation of the CFlips metric is that it does not measure the distance of each CF sample from the original data point. However, from an individual-fairness wise, a debated issue is the definition of a metric that considers that distance [5]. Accordingly, we propose a new metric that considers CFs ranked based on the Mean Absolute Deviation from the original sample and other criteria [8]. The insight behind this metric is that the more the CF is ranked high (in the top positions of the ranking), the more its impact on the metric value. Thus, the metric penalizes CFs ranked in the top positions for which the value of the sensitive feature is flipped. More formally:

Definition 3.3 (Discounted Cumulative Counterfactual Fairness)**.** Given a set of Counterfactuals $\mathbf{C_x}$ for a sample $\mathbf{x}_i$, the *Discounted Cumulative Counterfactual Fairness* $\text{DCCF}_{\mathbf{x}_i}$ measures the cumulative gain of the ranking of counterfactuals w.r.t. the sensitive group of the original sample:

$$\text{DCCF}_{\mathbf{x}_i} \triangleq \sum_{p_j, \mathbf{c}^j_{\mathbf{x}_i} \in C_{\mathbf{x}_i}} \frac{2^{(1-\mathbb{1}(c^j_{\mathbf{x}_i}))} - 1}{\log_2(p_j + 1)} \tag{4}$$

where p_j is the rank of $\mathbf{c}^j_{\mathbf{x}_i}$ in $C_{\mathbf{x}_i}$ and $\mathbb{1}(c^j_{\mathbf{x}_i})$ from Eq. 2.

If more CF samples belonging to the same sensitive group as the original data point are in a higher ranking position, the result will be a higher DCCF. Thereby, we can formulate the *Ideal Discounted Cumulative Counterfactual Fairness* (IDCCF) as an ideal ranking in which each CF sample $\mathbf{c_x}$ belongs to the same sensitive group as the original sample $\mathbf{x}$ (Eq. 5), and the *normalized* DCCF (nDCCF) (Eq. 6).

$$\text{IDCCF}_{\mathbf{x}_i} \triangleq \sum_{p_j, \mathbf{c}^j_{\mathbf{x}_i} \in C_{\mathbf{x}_i}} \frac{2^{(1)} - 1}{\log_2(p_j + 1)} \tag{5}$$

$$\text{nDCCF}_{\mathbf{x}_i} \triangleq \frac{\text{DCG}_{\mathbf{x}_i}}{\text{IDCG}_{\mathbf{x}_i}} \tag{6}$$

In the same way as CFlips, given a set of samples $\mathcal{X}^- \subseteq \mathcal{D}$ predicted by the decision model as negative, the metric in Eq. 6 can be generalized to the *unprivileged* and *privileged* group (Eq. 7).

$$\text{nDCCF}_s \triangleq \frac{1}{|\mathcal{X}|_s^-|} \sum_{\mathbf{x}_i} \text{nDCG}_{\mathbf{x}_i} \quad \text{with } \mathbf{x}_i \in \mathcal{X}|_s^- \tag{7}$$

For both CFlips and nDCCF, we are interested in the difference (i.e., Δ), between *privileged* and *unprivileged*, being close to zero.

4 EXPERIMENTAL EVALUATION

Dataset. The experimental evaluation has been carried out on three state-of-the-art benchmark datasets (i.e., Adult[2], Crime[2], and German[2]). We decided to create two different settings from the Adult dataset. The first (Adult) consists of the dataset itself, removing only the sensitive feature we take into account in the analysis (i.e., gender). In the second (Adult-debiased), we removed all the sensitive features (i.e., gender, age, marital status, and race), and all the non-sensitive features highly correlated with at least one sensitive feature (i.e., Pearson's correlation coefficient greater than 0.35). We do not include any sensitive features for training the model, guaranteeing the *fairness under unawareness* setting. Additional information is in Table 1.

Table 1: Overview of relevant dataset information, including sensitive feature distribution, target distribution, name of privileged group, and ex-ante Statistical Parity.

Dataset	$\lvert\mathcal{D}\rvert$	$\lvert n\rvert$	Y	$Y=1$	s	$s=1$	$\Phi(s)^\dagger$	$\Phi(Y)^{\dagger\dagger}$	ex-ante SP*
Adult	45222	13	income	$\geq \$50k$	gender	male	0,675/0,325	0,248/0,752	0.199
Adult-deb.	45222	6	income	$\geq \$50k$	gender	male	0,675/0,325	0,248/0,752	0.199
Crime	1994	98	Violent State	$<$*median*	race	white	0,58/0,42	0,5/0,5	0.554
German	1000	17	credit score	Good	gender	male	0,690/0,310	0,7/0,3	0.075

† Probability distribution of the *privileged* and *unprivileged* group: $\mathbb{P}(S=1)/\mathbb{P}(S=0)$

†† Probability distribution of the target variable: $\mathbb{P}(Y=1)/\mathbb{P}(Y=0)$

* A priori Statistical Parity probability: $\mathbb{P}(Y=1 \mid S=1) - \mathbb{P}(Y=1 \mid S=0)$

Table 2: XGB (i.e., $f_s(\cdot)$) results on the sensitive information.

Dataset	s	AUC	ACC	Recall	Precision	F1
Adult	gender	0.9411	0.8457	0.9634	0.8018	0.8752
Adult-deb	gender	0.7803	0.7404	0.8113	0.8022	0.8067
Crime	race	0.9896	0.9450	0.9411	0.9655	0.9532
German	gender	0.7139	0.6900	0.7159	0.9130	0.8025

Decision Maker. To keep the approach as general as possible, we opted for Logistic Regression[3] (LR), Support-Vector Machines[3] (SVM), XGBOOST[3] (XGB) , and LightGBM[3] (LGBM).

Debiased Decision Maker. To investigate the quality and the reliability of our metrics we used also two debiased classifiers, *Adversarial Debiasing*[3] (AdvDeb) proposed by Zhang et al. [12] and *Linear Fair Empirical Risk Minimization*[3] (lferm) proposed by Donini et al. [4] as in-processing algorithms.

Counterfactual Generator. For the sake of reproducibility and reliability, the counterfactuals are generated with an external counterfactual framework, DiCE [8], with $|C_{\mathbf{x}}|$ equal to 100[4].

Sensitive-Feature Classifier. We used XGB for implementing this component due to its capability to learn non-linear dependencies. Results split for dataset/sensitive feature are available in Table 2.

Metrics. We evaluate the models performance with the Area Under the Receiver Operative Curve (AUC), Accuracy (ACC), Recall, Precision, F1 score, and fairness by measuring Statistical Parity[5] (DSP), Equal Opportunity[6] (DEO), and Average Odds[7] (DAO).

Split and Hyperparameter Tuning. The datasets have been split with the hold-out method 90/10 train-test set, with stratified sampling w.r.t. the target and sensitive labels, to respect the original

[2] ADULT: https://archive.ics.uci.edu/ml/datasets/adult; CRIME: https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990); GERMAN:https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data). German results are reported in the repository.

[3] LR, SVM: https://scikit-learn.org/; XGB: https://github.com/dmlc/xgboost; LGBM: https://github.com/microsoft/LightGBM; AdvDeb: https://github.com/Trusted-AI/AIF360; lferm: https://github.com/jmikko/fair_ERM;

[4] DiCE offers several strategies for generating candidate counterfactual samples, but we choose to only exploit the Genetic one.

[5] $DSP = \left|\mathbb{P}(\hat{Y}=1 \mid S=1) - \mathbb{P}(\hat{Y}=1 \mid S=0)\right|$

[6] $DEO = \left|\mathbb{P}(\hat{Y}=1 \mid S=1, Y=1) - \mathbb{P}(\hat{Y}=1 \mid S=0, Y=1)\right|$

[7] $DAO = \frac{1}{2}\left|\sum_{Y\in\{0,1\}} (\mathbb{P}(\hat{Y}=1 \mid S=1, Y) - \mathbb{P}(\hat{Y}=1 \mid S=0, Y))\right|$

Table 3: AUC, ACC, DSP, DEO, and DAO results on Test set; CFlip and nDCG results at different $|k|$ number of Counterfactuals for each negatively predicted Test set sample. Due to space constraints, German is remanded to the repository .

							CFlips@$\|k\|$ (%)									nDCCF@$\|k\|$								
							Privileged			Unprivileged			ΔCFlips ↓			Privileged			Unprivileged			ΔnDCCF ↓		
Dataset	$f(\cdot)$	AUC↑	ACC↑	DSP↓	DEO↓	DAO↓	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100	@10	@50	@100
Adult	LR	0.9078	0.8099	0.2947	0.0546	0.1241	12.332	10.886	10.212	66.353	72.932	77.165	54.021	62.046	66.953	0.8678	0.8849	0.886	0.3522	0.2913	0.2497	0.5156	0.5936	0.6363
	SVM	0.9073	0.8541	0.1769	0.0644	0.0692	6.752	7.533	7.742	77.095	80.973	81.372	70.343	73.44	73.63	0.9306	0.9258	0.9171	0.2474	0.2042	0.1948	0.6832	0.7216	0.7223
	LGB	0.9304	0.8658	0.1850	0.0379	0.0569	9.195	8.541	8.781	65.918	76.605	79.697	56.723	68.064	70.916	0.9049	0.9124	0.9049	0.3611	0.2633	0.2272	0.5438	0.6491	0.6777
	XGB	**0.9314**	**0.8698**	0.1884	0.0635	0.0680	10.011	8.788	9.07	64.796	76.243	79.512	54.785	67.455	70.442	0.8968	0.9088	0.9014	0.3708	0.2677	0.2298	0.526	0.6411	0.6716
	AdvDeb	0.9123	0.8512	**0.1151**	0.1399	0.0879	30.046	34.488	34.968	36.11	38.694	43.041	6.064	**4.206**	**8.073**	0.7016	0.6668	0.6537	0.6427	0.6199	0.5812	0.0589	**0.0469**	**0.0725**
	lferm	0.9031	0.8428	0.1448	**0.0194**	**0.0386**	31.459	28.632	24.965	31.764	47.464	57.47	**0.305**	18.832	32.505	0.6857	0.7062	0.7314	0.6864	0.5632	0.4701	**0.0007**	0.143	0.2613
AdultDeb	LR	0.8233	0.7367	0.1567	0.0695	0.0693	8.438	10.838	13.192	54.816	57.521	57.047	46.378	46.683	43.855	0.9239	0.9012	0.8736	0.464	0.4332	0.4303	0.4599	0.468	0.4433
	SVM	0.8189	0.7395	0.1062	**0.0140**	**0.0152**	11.937	16.377	17.379	31.305	33.869	35.385	**19.368**	**17.492**	**18.006**	0.8871	0.8468	0.8295	0.6661	0.6616	0.6449	**0.221**	**0.1852**	**0.1846**
	LGB	**0.8596**	0.8371	0.1093	0.0470	0.0356	4.624	9.419	12.848	66.966	74.223	73.445	62.342	64.804	60.597	0.9578	0.9182	0.8815	0.3720	0.2863	0.2794	0.5858	0.6319	0.6021
	XGB	0.8578	**0.8375**	0.1056	0.0400	0.0304	1.803	3.152	6.523	81.289	88.9	84.48	79.486	85.748	77.957	0.9804	0.9711	0.9386	0.2183	0.1378	0.1599	0.7621	0.8333	0.7787
	AdvDeb	0.8309	0.8195	0.0957	0.0326	0.0282	17.041	20.686	23.588	44.315	52.371	56.786	27.274	31.685	33.198	0.8425	0.8055	0.7735	0.5852	0.5031	0.4566	0.2573	0.3024	0.3169
	lferm	0.8017	0.7953	**0.0639**	0.0179	0.0186	8.943	13.316	16.561	47.036	54.87	55.83	38.093	41.554	39.269	0.9248	0.8809	0.8452	0.5618	0.4791	0.4584	0.363	0.4018	0.3868
Crime	LR	0.9248	**0.8700**	0.6535	0.3294	0.3438	2.857	3.429	3.714	75.286	81.914	85.043	72.429	78.485	81.329	0.9688	0.9656	0.9564	0.2659	0.2015	0.1688	0.7029	0.7641	0.7876
	SVM	**0.9288**	**0.8700**	0.6395	0.3843	0.3390	6.667	5.917	5.671	73.38	80.676	84.437	66.713	74.759	78.766	0.9334	0.939	0.9349	0.2858	0.2157	0.1781	0.6476	0.7233	0.7568
	LGB	0.9168	0.8400	0.6363	0.2824	0.3525	5.455	5.818	5.636	74.571	80.229	83.693	69.116	74.411	78.057	0.9432	0.9417	0.9364	0.2875	0.2207	0.1842	0.6557	0.721	0.7522
	XGB	0.9099	0.8500	0.6568	0.2941	0.3656	4.762	5.429	5	73.38	80.113	83.712	68.618	74.684	78.712	0.9505	0.9469	0.943	0.2938	0.2216	0.1844	0.6567	0.7253	0.7586
	AdvDeb	0.9008	0.8100	**0.5501**	**0.1882**	**0.2732**	7.5	6.875	6.969	69	77.743	80.857	61.5	70.868	73.888	0.9302	0.931	0.9237	0.3396	0.2506	0.2146	0.5906	0.6804	0.7091
	lferm	0.9100	0.8400	0.6125	0.2941	0.3278	3.182	6	6.636	64.412	71.647	75.147	**61.23**	**65.647**	**68.511**	0.9679	0.9439	0.9306	0.3695	0.3045	0.2681	**0.5984**	**0.6394**	**0.6625**

distribution in each split. The Decision Maker, the Debiased models, and the Sensitive-Feature Classifier have been tuned on the training set with a Grid Search k-fold (k=5) cross-validation methodology, the first two optimizing AUC metric, and the latter F1 score to prevent unbalanced predictions on the sensitive feature.

5 RESULTS AND CONCLUSION

Table 3 summarizes the performance for each dataset and model. In most cases, debiased models (i.e., AdvDeb and lferm) have better performance in terms of fairness than other ML models. In all cases, Accuracy is sacrificed for fairness. Looking at our proposed fairness metrics which are measured for different k values (generated CF samples), we can see that *unprivileged* groups generally have more CFlips than *privileged* ones. This means that to achieve a favorable outcome, counterfactuals of *unprivileged* groups need to take on characteristics of the *privileged* samples. Similarly, nDCCF values for the *unprivileged* group has lower values than the *privileged* one. This means that counterfactuals of the *unprivileged* group in the highest positions of the ranking (i.e., most similar to the original sample) are classified by the Sensitive-Feature Classifier as *privileged* (opposite to the original class).

For the Adult dataset and the two debiased models (i.e., AdvDeb and lferm) the Δ is close to zero for both our metrics, meaning that there is not a great difference in the CFlips for both groups (privileged and unprivileged one). The debiased models perform the same both with standard fairness metrics and our metrics (i.e., CFlips, nDCCF). The same is not true for the debiased version of Adult dataset (i.e., AdultDeb) where debiased models turn out to perform better with the standard fairness metrics (i.e., DSP, DEO, DAO) than the new ones (i.e., CFlips, nDCCF). Indeed, SVM have the best ΔCFlips and ΔnDCCF performance but the worst accuracy. This is due to the difficulty of SVM to capture correlations between sensitive and non-sensitive features, and thus it learns a model that is fairer than others. In the case of CRIME, each model proves to be extremely biased with respect to the *privileged* class, with slightly better performance for the debiased models. Our metrics turn out to be consistent w.r.t. DSP, DEO, and DAO.

In conclusion, we present a novel methodology for detecting biases in decision-making models that do not use sensitive features and work in a context of fairness under unawareness. A new fairness concept (i.e., *Counterfactual Fair Opportunity*) and two related fairness metrics (i.e., CFlips and nDCCF) are proposed. Understanding how an algorithm can behave with new samples and how the traits of favored groups can influence a favorable result is crucial. In the case of sensitive feature blindness, counterfactual reasoning, and, more specifically, the methodology proposed in this paper can be an effective tool for confirming and assessing the discriminatory behavior of ML models.

REFERENCES

[1] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *FAT*. ACM, 339–348.

[2] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *KDD*. ACM, 797–806.

[3] Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, and Eugenio Di Sciascio. 2023. Auditing fairness under unawareness through counterfactual reasoning. *Information Processing & Management* 60, 2 (2023), 103224. https://doi.org/10.1016/j.ipm.2022.103224

[4] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical Risk Minimization Under Fairness Constraints. In *NeurIPS*. 2796–2806.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *ITCS*. ACM, 214–226.

[6] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS (LIPIcs, Vol. 67)*.

[7] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *NIPS*. 4066–4076.

[8] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.

[9] Saerom Park, Seongmin Kim, and Yeon-sup Lim. 2022. Fairness Audit of Machine Learning Models with Confidential Computing. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 3488–3499.

[10] Judea Pearl. 1994. Causation, Action and Counterfactuals. In *ECAI*. John Wiley and Sons, Chichester, 826–828.

[11] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[12] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AIES*. ACM, 335–340.