Ensuring Fairness with Transparent Auditing of Quantitative Bias in AI Systems

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—With the rapid advancement of AI, there is a growing trend to integrate AI into decision-making processes. However, AI systems may exhibit biases that lead decision makers to draw unfair conclusions. Notably, the COMPAS system used in the American justice system to evaluate recidivism was found to favor racial majority groups; specifically, it violates a fairness standard called equalized odds. Various measures have been proposed to assess AI fairness. We present a framework for auditing AI fairness, involving third-party auditors and AI system provider, and we have created a tool to facilitate systematic examination of AI systems. Inclusion of independent and trusted auditors is pivotal for objectivity and accountability; third parties are often further necessary for their specialized expertise in relevant domains. Auditors, equipped with our tool, can thoroughly review AI systems for bias and fairness violations. The tool is open-sourced and publicly available. Unlike traditional AI systems, we advocate a transparent white-box and statistics-based approach. It can be utilized by third party auditors, AI developers, or the general public for reference when judging the fairness criterion of AI systems.**

***Index Terms*—AI, fairness, auditing**

I. Introduction

In recent years, the accelerating pace of artificial intelligence (AI) technologies have revolutionized numerous fields. In healthcare, AI-driven diagnostic systems streamline disease identification, while in finance, automated trading algorithms analyze market trends to execute optimal trades swiftly. This remarkable progress has reached diverse industries, sprouting out various applications for different purposes.

One of its most profound impact is how AI has transformed decision-making processes across sectors. By harnessing vast amounts of data and employing advanced algorithms, AI empowers organizations to make more informed and strategic decisions. From assessing job applicants to determining school admissions, AI-driven insights offer efficient and analytical advantages.

AI system, however, may be biased. Factors such as inherent biases in original data sets or flaws in algorithm designs could contribute to bias within AI systems. When left unchecked, the ramifications of such biases extend far beyond mere inaccuracies; it can lead to devastating consequences, such as amplifying systemic injustices, perpetuating group discrimination, and exacerbating societal inequalities.

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a tool for predicting recidivism—the tendency of criminals to reoffend. It is used in the criminal justice system in multiple states in the United States. In 2016, it was discovered in an investigation by the journalists at ProPublica that the COMPAS system is, in fact, unfair towards minority and disadvantaged groups.

Cases such as COMPAS underscore the critical importance of rigorously examining the fairness of AI systems by third parties. Fairness is fundementally a subjective social construct, heavily influenced by cultural context and deeply rooted in historical inequalities. However, there have been development in research leveraging statistical metrics to quantify fairness that provide transparent and objective insights.

By employing such mathematically rigorous methodologies, our research can offer a more deeper look into the fairness of AI systems, enabling third parties to make more informed and robust judgment. For example, it would reveal that the COMPAS system violates a fairness measure called equalized odds.

In addition to equalized odds, there are a number of fairness measures for AI systems. Our framework aims to help third parties apply these metrics and navigate complex datasets with ease, pinpointing instances of bias and discrimination.

With our tool, a third party such as ProPublica could clearly and easily demonstrates that COMPAS violates equalized odds according to their datasets. Our platform empowers organizations to conduct thorough assessments of AI fairness and uncover discrepancies in predictive models.

This functionality not only facilitates sound judgment but also fosters objectivity and transparency regarding AI systems, ultimately bolster the claim of fairness and justice in AI-enabled decision-making.

Our tool is written in Python and is offered as a Python package. It supports common datasets format such as csv. It is open-sourced and publicly available for downloads.

II. Fairness Measures

Fairness is about making sure the disadvantaged and unprivileged groups of individuals are treated equitably. However, there have been several interpretations of it proposed in the past. To have a constructive discussion on fairness, we

Identify applicable funding agency here. If none, delete this.

must first have precise definitions of it. Pessach and Shmueli [1] formulated a number of fairness measures in an unified mathematical notation. We will base our framework on their formulation.

A. Preliminaries

For the following definition we will use Y to denote the ground truth of an outcome; $\hat{Y}$ to denote the predicated result of an outcome; $Y = 1$ and $\hat{Y} = 1$ to denote them being accepted or positive. For example, let Y be recidivism. Then $Y = 1$ means the case of an individual actually recidivating and $\hat{Y} = 0$ means the predication of an individual recidivating is negative. In addition, we use V and $\hat{V}$ when the truth and the prediction isn't binary. For example, we denote COMPAS score by $\hat{V}$, which ranges from 1 to 10.

We denote by S some protected attribute. Protected attributes are the characteristics of individuals that are, for example, legally or ethically, considered sensitive and warrant protection against discrimination and bias. We write $S = 1$ to represent the privileged group and $S \neq 1$ to represent the unprivileged group. For example, let S be Caucasian. Then $S = 1$ represents the case of an individual's race being Caucasian and $S \neq 1$ vice versa.

We denote by ϵ some threshold that's used to limit the fairness measures.

B. Disparate Impact

In 1971 [2], the US spreme court ruled that it is illegal for hiring decisions to have "disparate impact" by race, thus coining the term. It is taken as unintentional discrimination, as opposed to intentional discrimination, which is called "disparate treatment".

Legal cases involving disparate impact often refer to the "80% Rule", advocated by the US Equal Employment Opportunity Commission [3], where it requires the selection rate of a minority group is to be no less than 80% of that of a majority group. Formally [4], it is:

$$\frac{P[\hat{Y} = 1 | S \neq 1]}{P[\hat{Y} = 1 | S = 1]} \geq 1 - \epsilon$$

In the case of 80% rule, $\epsilon = 20\%$.

C. Demographic Parity

Demographic parity [5], also known as statistical parity, is similar to disparate impact but, instead of ratio, difference is taken. It is named so to suggest that each demographic group(such as race, gender, or age) should have equal representation or opportunity. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1] - P[\hat{Y} = 1 | S \neq 1]| \leq \epsilon$$

D. Conditional Statistical Parity

Conditional statistical parity [6] is similar to demographic parity, but, in addition to protected attributes, it further takes into account some "legitimate" attributes that are *legitimately* related to the case. For example, a legitimate attribute when considering future recidivism could be the number of prior crimes committed. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1, L = l] - P[\hat{Y} = 1 | S \neq 1, L = l]| \leq \epsilon$$

where L denotes the legitimate attributes.

E. Overall Accuracy Equality

Overall accuracy eqaulity [7] is similar to demographic parity, but instead of the case of $\hat{Y} = 1$, it considers the case of $Y = \hat{Y}$; that is, the case where the prediction is accurate. Formally, it is:

$$|P[Y = \hat{Y} | S = 1] - P[Y = \hat{Y} | S \neq 1]| \leq \epsilon$$

F. Mean Difference

Mean difference [8] considers the case of non-binary outcome, such as the score of COMPAS. It takes the expected value of the prediction. Formally, it is:

$$|E[\hat{Y} | S = 1] - E[\hat{Y} | S \neq 1]| \leq \epsilon$$

G. Equalized Odds

Equalized odds [9] is similar to demogrphic parity, but it further takes into account the grouth truth. It consider the cases of true positive and false positive. It solves the downsides of a fully accurate classifier might be deemed unfair by measures that don't take ground truth into consideration. For example, consider a group A that is predicated to recidivate and they do in fact always recidivate and a group B that is predicated to recidivate but end up never recidivating. A fully accurate classifier will always predict A to recidivate while B to never recidivate, and this will violate demogrphic parity. By taking ground truth into account, equalized odds avoids these pitfalls. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1, Y = 0] - P[\hat{Y} = 1 | S \neq 1, Y = 0]| \leq \epsilon$$
$$|P[\hat{Y} = 1 | S = 1, Y = 1] - P[\hat{Y} = 1 | S \neq 1, Y = 1]| \leq \epsilon$$

H. Equal Opportunity

Equal opportunity [9] is a relaxation of equalized odds by only considering the true positive case. It is named so because when outcome being positive is beneficial to the individual, as it is an opportunity, such as school acceptance, the fairness of the true positive case is much more important. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1, Y = 1] - P[\hat{Y} = 1 | S \neq 1, Y = 1]| \leq \epsilon$$

I. Predictive Equality

Predictive equality [6] is also a relaxation of equalized odds by only considering the false positive case. It was formulated by considering positive outcome being detrimental to the individual, such as predicting future recidivism. Thus, only the false positive case is taken. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1, Y = 0] - P[\hat{Y} = 1 | S \neq 1, Y = 0]| \leq \epsilon$$

J. Conditional Use Accuracy Equality

Conditional use accuracy equality [7] is similar to equalized odds, but instead of conditioning on the ground truth, it conditions on the prediction and calculated the probability of the ground truth. It can be seem as checking the prediction accuracy across groups, thus the name. It further requires the measure on the case of positive predictive values is less than that of the negative predictive values. Formally, it is:

$$\begin{aligned} |P[Y=1|S=1,\hat{Y}=1]-P[Y=1|S\neq 1,\hat{Y}=1]| &\leq \epsilon \\ \wedge \\ |P[Y=0|S=1,\hat{Y}=0]-P[Y=0|S\neq 1,\hat{Y}=0]| &\leq \epsilon \end{aligned}$$

K. Predictive Parity

Predictive parity [10] is a relaxation of conditional use accuracy equality by only considering the positive predictive value case. Like predictive equality, it was formulated by considering positive outcome being detrimental to the individual. Thus, only the positive predictive value case is taken. Formally, it is:

$$|P[Y=1|S=1,\hat{Y}=1]-P[Y=1|S\neq 1,\hat{Y}=1]| \leq \epsilon$$

L. Equal Calibration

Equal calibration [10] is similar to equal opportunity, but instead of having a binary $\hat{Y}$, it is conditioned on the range of the predicted value $\hat{V}$. For example, this could be conditioned on the highest COMPAS score $\hat{V} = 10$. Calibration is a concept of having fair score function [11]. Formally, it is:

$$|P[Y=1|S=1,\hat{V}=v]-P[Y=1|S\neq 1,\hat{V}=v]| \leq \epsilon$$

M. Positive Balance

Positive balance [12] is similar to equal opportunity, but instead of taking the difference of probability of binary prediction $\hat{Y}$, it takes the difference of the expected value of the score $\hat{V}$. Formally, it is:

$$|E[\hat{V}|Y=1,S=1]-E[\hat{V}|Y=1,S\neq 1]| \leq \epsilon$$

N. Negative Balance

Negative balance [12] is like positive balance except it conditions on the case of $Y = 0$. Formally, it is:

$$|E[\hat{V}|Y=0,S=1]-E[\hat{V}|Y=0,S\neq 1]| \leq \epsilon$$

These fairness measures are compiled in Table I.

III. Auditing Framework

Per the review by Pessach and Shmueli [1], we designed an auditing framework for calculating the various fairness measure. We offer two versions of fairness checkers: one for when the prediction positive results are readily available in csv input and one for when a model is provided. In most auditing cases the model version is preferred because csv result can be easily fabricated.

A. Preliminaries

Let $\mathcal{D} = \{r_1, r_2, ...\}$ be a database containing rows of data. Each row $r_i : \texttt{str} \rightarrow \texttt{str}$ is a lookup table or dictionary. For example, r_n("sex") = "Female" means r_n's sex is female.

A model $\mathcal{M} : (\texttt{str} \rightarrow \texttt{str}) \rightarrow \alpha$ takes a row and returns the model's prediction result of some type α.

A privileged predicate $P : (\texttt{str} \rightarrow \texttt{str}) \rightarrow \texttt{bool}$ takes a row and determines if it belongs to the privileged group. For example, $P(r_i) := r_i$("race") == "Caucasian" means the privileged group is those with race being Caucasian.

A positive predicate of rows $\hat{P} : (\texttt{str} \rightarrow \texttt{str}) \rightarrow \texttt{bool}$ takes a row and determines if its prediction is positive. For example, $\hat{P}(r_i) := \texttt{int}(r_i(\text{"score"})) >$ "7" means a row's prediction is positive if its score is greater than 7. A positive predicate of model results $\hat{P} : \alpha \rightarrow \texttt{bool}$ takes a model's result and determines if it is positive. In the simplest case, the model returns a `bool`, and the positive predicate can just be the idnetity function.

A ground truth predicate $T : (\texttt{str} \rightarrow \texttt{str}) \rightarrow \texttt{bool}$ takes a row and gives the ground truth of the result.

A legitimate predicate $L : \alpha_i^n \rightarrow (\texttt{str} \rightarrow \texttt{str}) \rightarrow \texttt{bool}$, and similarly a score predicate $\hat{S} : \alpha_i^n \rightarrow (\texttt{str} \rightarrow \texttt{str}) \rightarrow \texttt{bool}$, takes n parameters and returns a row predicate. For example, $\hat{S}(u,l)(r_i) := l < \texttt{int}(r_i(\text{"score"})) < u$ first takes two parameters u, l as upper bound and lower bound, and then decides if r_i's prediction score is between them.

B. Definitions

We abstracted the idea of privileged groups and positive prediction as predicates to maximizes flexibility. With the proper predicates, any model output can be used; even prose-like response of generative models can be included given adequate predicates.

Given a datasets for auditing use, and given csv prediction results or the prediction model itself, we can calculate the fairness measures mentioned in Section II by:

$$\begin{aligned} \text{disparate\_impact}&(\epsilon, \mathcal{M}, P, \hat{P}) \\ \text{demographic\_parity}&(\epsilon, \mathcal{M}, P, \hat{P}) \\ \text{conditional\_statistical\_parity}&(\epsilon, \mathcal{M}, P, \hat{P}, L, (args, ...)) \\ \text{overall\_accuracy\_eqaulity}&(\epsilon, \mathcal{M}, P, \hat{P}, T) \\ \text{mean\_difference}&(\epsilon, \mathcal{M}, P, \hat{P}) \\ \text{equalized\_odds}&(\epsilon, \mathcal{M}, P, \hat{P}, T) \\ \text{equal\_opportunity}&(\epsilon, \mathcal{M}, P, \hat{P}, T) \\ \text{predictive\_equality}&(\epsilon, \mathcal{M}, P, \hat{P}, T) \\ \text{conditional\_use\_accuracy\_equality}&(\epsilon, \mathcal{M}, P, \hat{P}, T) \\ \text{predictive\_parity}&(\epsilon, \mathcal{M}, P, \hat{P}, T) \\ \text{equal\_calibration}&(\epsilon, \mathcal{M}, P, \hat{S}, T, (args, ...)) \\ \text{positive\_balance}&(\epsilon, \mathcal{M}, P, \hat{S}, T, (args, ...)) \\ \text{nagative\_balance}&(\epsilon, \mathcal{M}, P, \hat{S}, T, (args, ...)) \end{aligned}$$

Fairness Measure	**Definition**
Disparate Impact	$\frac{P[\hat{Y}=1\vert S\neq 1]}{P[\hat{Y}=1\vert S=1]} \geq 1-\epsilon$
Demographic Parity	$\vert P[\hat{Y}=1\vert S=1]-P[\hat{Y}=1\vert S\neq 1]\vert \leq \epsilon$
Conditional Statistical Parity	$\vert P[\hat{Y}=1\vert S=1,L=l]-P[\hat{Y}=1\vert S\neq 1,L=l]\vert \leq \epsilon$
Overall Accuracy Eqaulity	$\vert P[Y=\hat{Y}\vert S=1]-P[Y=\hat{Y}\vert S\neq 1]\vert \leq \epsilon$
Mean Difference	$\vert E[\hat{Y}\vert S=1]-E[\hat{Y}\vert S\neq 1]\vert \leq \epsilon$
Equalized Odds	$\vert P[\hat{Y}=1\vert S=1,Y=0]-P[\hat{Y}=1\vert S\neq 1,Y=0]\vert \leq \epsilon$
	$\vert P[\hat{Y}=1\vert S=1,Y=1]-P[\hat{Y}=1\vert S\neq 1,Y=1]\vert \leq \epsilon$
Equal Opportunity	$\vert P[\hat{Y}=1\vert S=1,Y=1]-P[\hat{Y}=1\vert S\neq 1,Y=1]\vert \leq \epsilon$
Predictive Equality	$\vert P[\hat{Y}=1\vert S=1,Y=0]-P[\hat{Y}=1\vert S\neq 1,Y=0]\vert \leq \epsilon$
Conditional Use Accuracy Equality	$\vert P[Y=1\vert S=1,\hat{Y}=1]-P[Y=1\vert S\neq 1,\hat{Y}=1]\vert \leq \epsilon$
	$\vert P[Y=0\vert S=1,\hat{Y}=0]-P[Y=0\vert S\neq 1,\hat{Y}=0]\vert \leq \epsilon$
Predictive Parity	$\vert P[Y=1\vert S=1,\hat{Y}=1]-P[Y=1\vert S\neq 1,\hat{Y}=1]\vert \leq \epsilon$
Equal Calibration	$\vert P[Y=1\vert S=1,\hat{V}=v]-P[Y=1\vert S\neq 1,\hat{V}=v]\vert \leq \epsilon$
Positive Balance	$\vert E[\hat{V}\vert Y=1,S=1]-E[\hat{V}\vert Y=1,S\neq 1]\vert \leq \epsilon$
Negative Balance	$\vert E[\hat{V}\vert Y=0,S=1]-E[\hat{V}\vert Y=0,S\neq 1]\vert \leq \epsilon$

TABLE I: Fairness measures.

where $\mathcal{M}$ is optional if the input contains csv prediction results.

By calculating all of the available fairness measures, we provide the auditors a more comprehensive perespective on a model or dataset.

IV. Application

In this section, we will apply the framework proposed in Section III to the ProPublica COMPAS dataset.

A. ProPublica COMPAS Dataset

propublica

References

[1] D. Pessach and E. Shmueli, "A review on fairness in machine learning," ACM Computing Surveys (CSUR), vol. 55, no. 3, pp. 1–44, 2022.

[2] Supreme Court of the United States, "Griggs v. duke power co." 401 U.S. 424, March 8, 1971.

[3] The U.S. Equal Employment Opportunity Commission (EEOC), "Uniform guidelines on employee selection procedures," March 2, 1979.

[4] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[5] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," Data mining and knowledge discovery, vol. 21, pp. 277–292, 2010.

[6] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, 2017, pp. 797–806.

[7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," Sociological Methods & Research, vol. 50, no. 1, pp. 3–44, 2021.

[8] I. Žliobaitė, "Measuring discrimination in algorithmic decision making," Data Mining and Knowledge Discovery, vol. 31, no. 4, pp. 1060–1089, 2017.

[9] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," Advances in neural information processing systems, vol. 29, 2016.

[10] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," Big data, vol. 5, no. 2, pp. 153–163, 2017.

[11] A. Fraenkel, Fairness and Algorithmic Decision Making, 2020, lecture Notes for UCSD course DSC 167.

[12] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," arXiv preprint arXiv:1609.05807, 2016.