MDPI

Article

Does the COMPAS Needle Always Point Towards Equity? Finding Fairness in the COMPAS Risk Assessment Algorithm: A Case Study

Amrita Acharya[1], Dianne Caravela[1], Eunice Kim[1], Emma Kornberg[1], Elisabeth Nesmith[1]

[1] Statistical and Data Sciences Smith College Northampton, MA 01063; bbaumer@smith.edu
* Correspondence: aacharya@smith.edu, dcaravela@smith.edu, ekim89@smith.edu, ekornberg@smith.edu, enesmith@smith.edu

Version May 2, 2022 submitted to Water

Abstract: A variety of disciplines use risk assessment instruments to help humans make data-driven decisions. Northpointe, a software company, created an algorithmic risk assessment instrument known as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). COMPAS uses various behavioral and psychological metrics related to recidivism to assist justice systems in assessing a defendant's potential recidivism risk. Angwin *et al.* [1] published a ProPublica article in which they conclude that the racial biases in the criminal justice system are reflected in the COMPAS recidivism risk scores. In response, Dieterich *et al.* [2] published a rebuttal on behalf of Northpointe defending the COMPAS algorithm and refuting Angwin *et al.* [1]'s allegation of racial bias. Using a human rights framework adopted from the organizations Women at the Table and AI Fairness 360, we use debiasing algorithms and fairness metrics to analyze the argument between Northpointe and ProPublica and determine whether and to what extent there is racial bias in the COMPAS algorithm. All four group fairness metrics determine that the COMPAS algorithm favors white defendants over Black defendants.

1. Introduction

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm was created by the private, for-profit company Northpointe (now known by its parent company equivant), to predict defendants' risk of recidivism. It generates a decile score that classifies defendants' risk of recidivism as either low, medium, or high [1]. Jurisdictions across the United States use the COMPAS risk assessment instrument, including but not limited to the New York, Massachusetts, Michigan, California, and Wisconsin Departments of Corrections.

Due to the proprietary nature of the COMPAS algorithm, it is unknown how exactly these recidivism risk scores are calculated. However, a sample COMPAS Risk Assessment Survey has been made publicly available, revealing the algorithm's input information. Angwin *et al.* [1] critiques this survey for using proxy variables for race that do not explicitly factor in a defendant's race but heavily imply it, allowing Northpointe to claim that their algorithm is free of racial bias. For example, the COMPAS risk assessment survey asks screeners to speculate if a defendant might be affiliated with a gang. It also asks if a defendant has any friends or family members who have been crime victims [**?**]. Although these questions do not directly ask about race, they do not take into account the pervasive nature of systemic racism that infiltrates every aspect of the lives of marginalized people, thereby indirectly asking about race.

Angwin *et al.* [1] analyzes the methods and algorithms used by Northpointe in their COMPAS risk score assessment algorithm and uncovers racial biases in defendants' scores [1]. They find that "the algorithm [is] somewhat more accurate than a coin flip," a worrisome level of accuracy given

the potential impact its determinations may have on real people's lives. Angwin *et al.* [1] specifically investigate the distribution of COMPAS scores by decile among Black and white defendants. They write: "The analysis also [shows] that even when controlling for prior crimes, future recidivism, age, and gender, Black defendants [are] 45 percent more likely to be assigned higher risk scores than white defendants" [3]. After examining the fairness metric statistical parity difference, Angwin *et al.* [1] conclude that the algorithm is racially biased [3].

Dieterich *et al.* [2], on behalf of Northpointe, deny the allegations of racial bias and offer their own analyses based on different fairness metrics in rebuttal [2]. Angwin *et al.* [1] maintain that there are biases in the outcome values, protected attributes, and covariates during Dieterich *et al.* [2]'s data processing phase. ProPublica collaborators Larson *et al.* [3] account for these biases in their analyses. In their response, Dieterich *et al.* [2] highlight that Angwin *et al.* [1] did not account for base rates of recidivism in their analysis, which are important initial percentages without the presence of other information.

Women at the Table, the sponsor organization for this project, is "a growing, global gender equality & democracy CSO based in Geneva, Switzerland focused on advancing feminist systems change by using the prism of technology, innovation & AI exercising leverage points in technology, the economy, sustainability & democratic governance." We are collaborating with the organization on its AI & Equality [4] initiative, tasked with de-biasing the COMPAS algorithm [5] and producing a corresponding data story that will be added to its library.

Our project builds on Women at the Table's various de-biasing algorithms used in its AI & Equality Human Rights Toolbox to conduct our own analyses on the COMPAS data set. Based on this analysis, we employ a human rights framework to contribute to the ProPublica and Northpointe debate and investigate whether or to what extent there is racial bias in the COMPAS algorithm. With a solid understanding of the two sides, we aim to pinpoint the shortcomings of both arguments and correct them in our analyses. We will use various de-biasing techniques and fairness metrics to evaluate the level of bias present in the COMPAS data and our algorithm. We will summarize our results using the JupyterNotebook framework from Women at the Table, to be used by members of the organization to teach in a workshop setting. We hope that our findings will highlight the importance of checking statistical analyses using varied methods and contribute to the ongoing discussion of the effects of machine biases in the justice system.

2. Data

The data we are using for this project is the COMPAS General Recidivism Risk Scores data set from the AI Fairness 360 (AIF360) toolkit [5], which does the same initial pre-processing as ProPublica. The raw data has 6,167 rows and each row represents an arrest charge for a defendant. AIF360's COMPAS data includes the defendant's age, race, sex, prior charges, what they were charged with, and whether or not the defendant ultimately recidivated within a two-year period after their arrest. For the purposes of our project, which endeavors to evaluate the differing effects of the COMPAS algorithm on white defendants versus Black defendants, we have filtered the data to only include individuals whose race is listed as Caucasian or African-American. Our data therefore has 5,273 rows (Figure 1), with the distributions of age (Figure 2), prior charges (Figure 3), and two-year recidivism by race (Figure 4) and (Table 1) shown below.

Table 1. Incidence of recidivism by race, illustrating how a much greater proportion (> 50%) of Black defendants recidivated than their white counterparts.

Two Year Recidivism by Race	Recidivated	Survived	Total
African American	1661	1512	3173
Caucasian	822	1278	2100
Total	2483	2790	5273

id	sex	sex	age	age_cat	race	juv_fel_count	juv_misd_count	juv_other_count	priors_count	c_charge_degree	c_charge_desc	two_year_recid
1	0	Male	34	25 - 45	African-American	0	0	0	0	F	Felony Battery w/Prior Convict	Recidivated
2	0	Male	24	Less than 25	African-American	0	0	1	4	F	Possession of Cocaine	Recidivated
4	0	Male	41	25 - 45	Caucasian	0	0	0	14	F	Possession Burglary Tools	Recidivated
6	1	Female	39	25 - 45	Caucasian	0	0	0	0	M	Battery	Survived
7	0	Male	27	25 - 45	Caucasian	0	0	0	0	F	Poss 3,4 MDMA (Ecstasy)	Survived

Figure 1. A snippet of the data set we will be using, containing information on a defendant's age, sex, race, criminal history, charge degree, charge description, and two-year recidivism outcome.

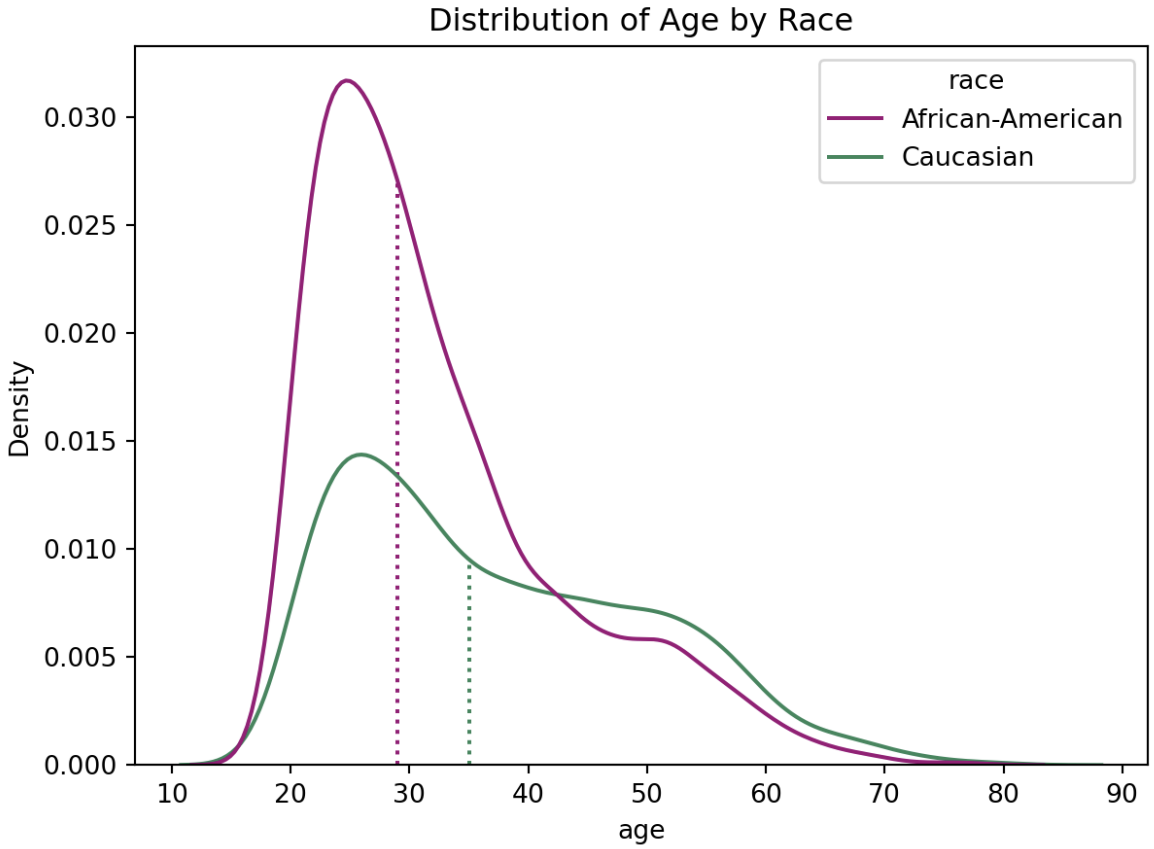

Figure 2. The purple curve shows the distribution of the ages of Black defendants, and the green curve shows the distribution of the ages of white defendants. The probability of a defendant's age being between two points on the x-axis is the total shaded area of the curve under the two points. The purple dotted line represents the median age of Black defendants (29 years) and the green dotted line represents the median age of white defendants (35 years). For both groups, the majority of defendants are relatively young, but this is especially noticeable for Black defendants.

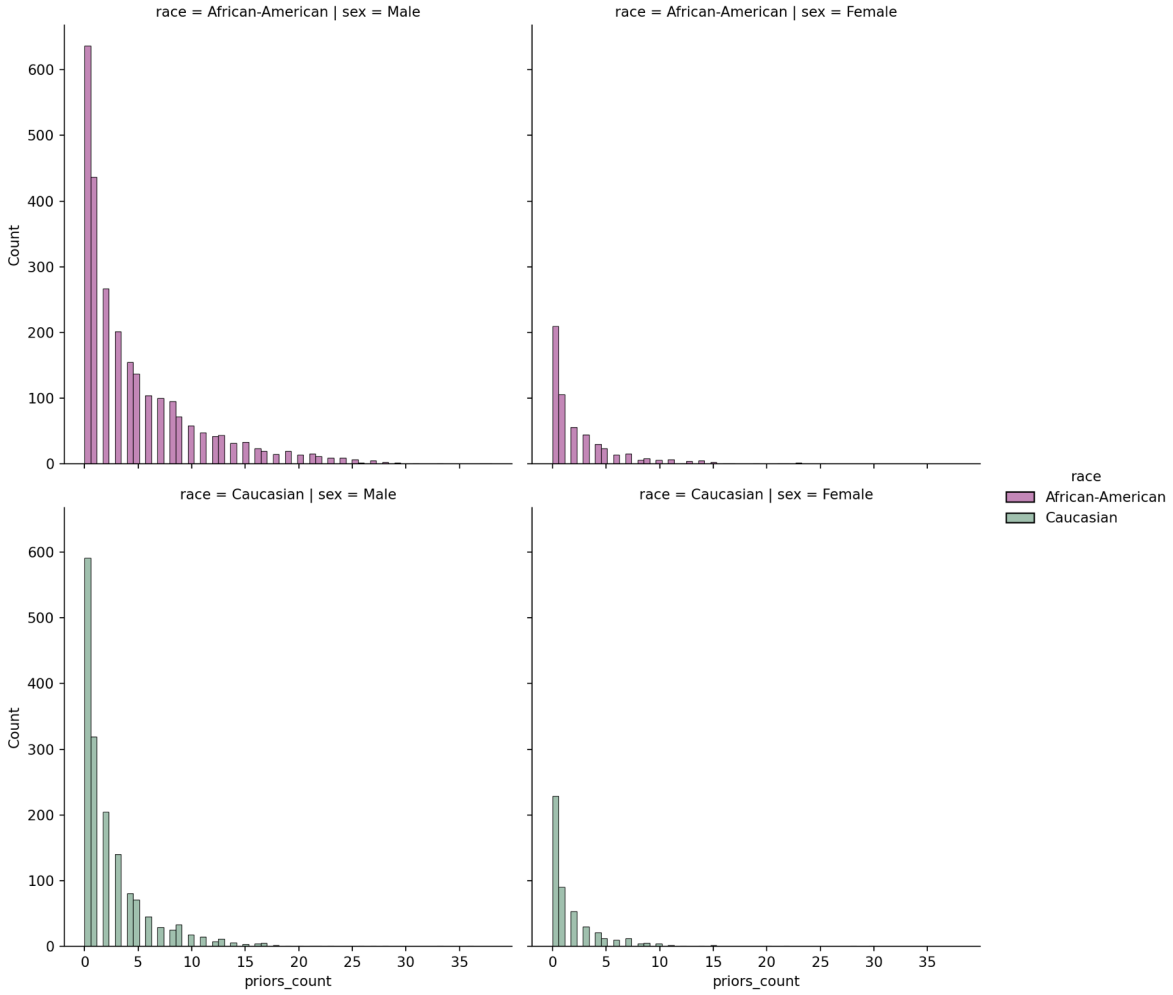

Figure 3. Black defendants, particularly men, are more likely to have a greater count of prior charges than white defendants. Male defendants have a higher number of prior charges than do female defendants. Though we do not know for sure which information goes into the COMPAS algorithm, it is likely that a defendant with prior charges will be coded as a having a higher risk of recidivism. Thus, by looking at the racial discrepancies in prior charges we can already see potential bias in the algorithm.

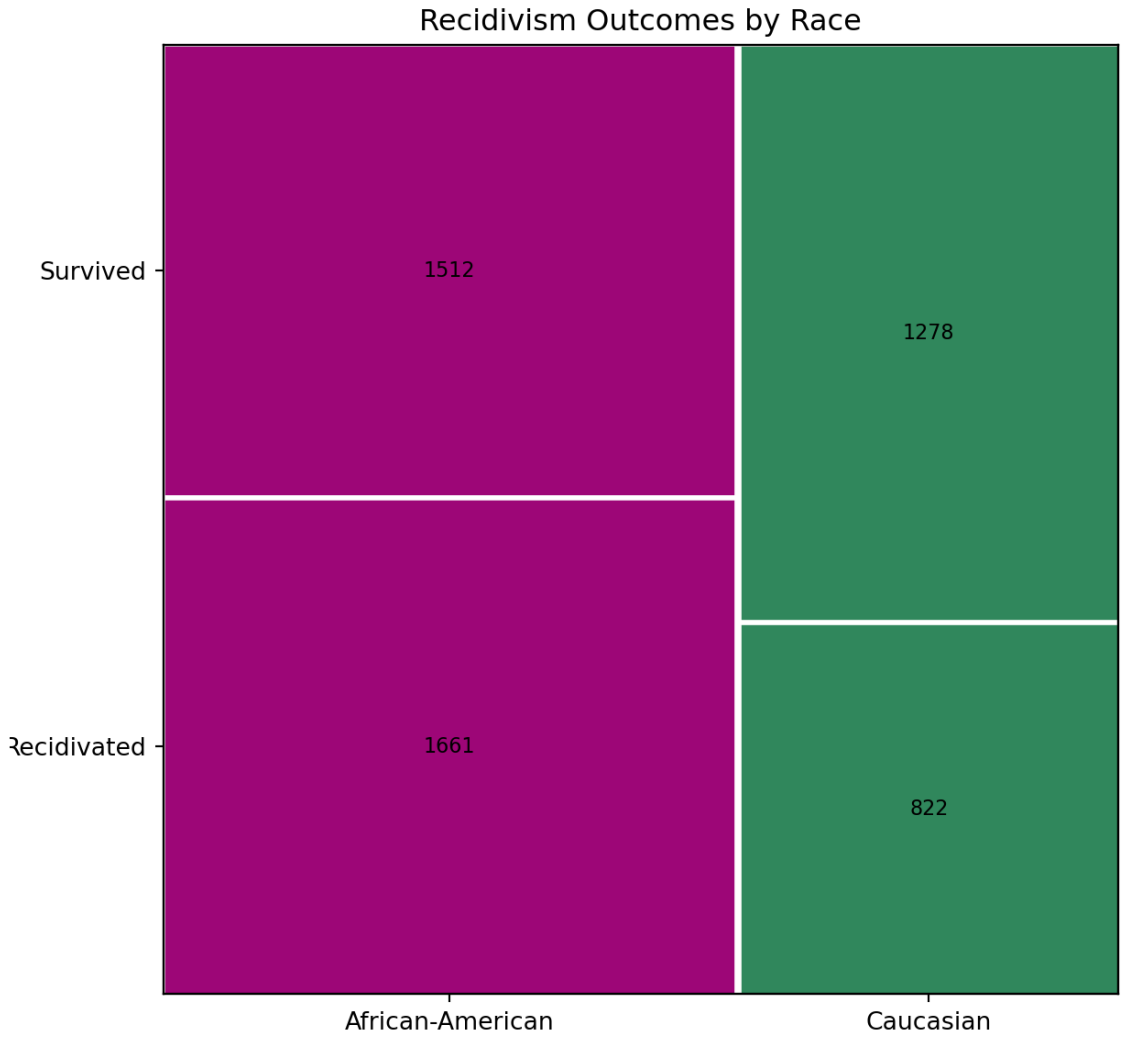

Figure 4. When we divide the data into Black and white defendants, we can see that Black defendants recidivate more than white defendants and Black defendants are more likely to recidivate than not recidivate. 39.14% of white defendants did recidivate within two years compared to 52.35% of Black defendants. We can also see that there are more Black defendants in the dataset overall.

3. Methods

AI Fairness 360 (AIF360) is an open-source Python toolkit that seeks "to help facilitate the transition of fairness research algorithms to use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms" [5]. It contains multiple data sets, including the COMPAS data set that accompanied Angwin *et al.* [1].

The AIF360 toolkit contains various group and individual fairness metrics as well as pre-processing, in-processing, and post-processing algorithms that we used to debias the COMPAS algorithm [5]. We researched the definitions and applications of different fairness metrics [6] to determine which metric would be most appropriate for our project. We chose to look at four group fairness metrics instead of individual fairness metrics. Group fairness metrics take into account the attributes of a whole group as opposed to just one individual in the group, allowing us to represent systemic issues. In general, group fairness metrics require that the unprivileged group is treated similarly to the privileged group, whereas individual fairness metrics require individuals to be treated consistently [7]. Group and individual metrics work in opposition of one another, meaning that when group fairness improves, individual fairness gets worse [7].

3.1. Statistical Parity Difference

This metric measures the difference between privileged and marginalized groups' likelihood to get a particular outcome. The ideal value of this metric is 0. Fairness for this metric is between -0.1 and 0.1. A negative value means there is higher benefit for the privileged group (in this case, white defendants).

$$P(\hat{Y} = 1|D = Unprivileged) - P(\hat{Y} = 1|D = Privileged)$$

3.2. Disparate Impact Ratio

This metric is the ratio of how often the favorable outcome occurs in one group versus the other. In the case of recidivism, this is the ratio of how many white defendants are predicted to not recidivate compared to how many black defendants are predicted to not recidivate. A value of 1 means that the ratio is exactly 1:1. Less than 1 means the privileged group (white defendants) benefits, while a value greater than 1 means the unprivileged group (Black defendants) benefits. According to AIF360, a ratio between 0.8 to 1.25 is considered fair [8].

$$\frac{P(\hat{Y} = 1|D = Unprivileged)}{P(\hat{Y} = 1|D = Privileged)}$$

3.3. Equal Opportunity Difference

This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group [**?**].

The ideal value is 0. A value less than 0 implies higher benefit for the privileged group and a value greater than 0 implies higher benefit for the unprivileged group. Fairness for this metric is between -0.1 and 0.1 [5].

This metric is best used when it is very important to catch positive outcomes while false positives are not exceptionally problematic [9]. This is not the case for the COMPAS data set, as false positives mean extra jail time for someone who will not actually re-offend.

$$TPR_{D=Unprivileged} - TPR_{D=Privileged}$$

3.4. Average Odds Difference

This metric returns the average difference in false positive rate and true positive rate for the privileged and unprivileged groups. A value of 0 indicates equality of odds, and a value below 0 implies benefit for the privileged group. Equality of odds is achieved in the case of recidivism when the proportion of people who were predicted to recidivate and did recidivate is equal (true positive rate) for both Black and white defendants AND the proportion of people who were predicted to recidivate and did not recidivate (false positive rate) is equal for both Black and white defendants [5].

$$\frac{1}{2}\left[(FPR_{D=Unprivileged} - FPR_{D=Privileged}) + \underbrace{(TPR_{D=Unprivileged} - TPR_{D=Privileged})}_{\text{Equal Opportunity Difference}}\right]$$

For the next step of our experiment, we need to determine where in the data science pipeline we can mitigate the most bias, using pre-processing, in-processing, and post-processing de-biasing algorithms. These are all based on using predictive models to figure out how we can "fix" the bias that is present.

Pre-processing refers to mitigating bias within the training data, and it is the most flexible method because it has not yet trained a model that may carry assumptions about the data. It is important to keep in mind that pre-processing prevents assumptions in the modeling, but does not account for the bias in data collection. Training data is where bias is most likely to be introduced. We use the reweighing pre-processing algorithm from AIF360 which assigns weights to the data. "The advantage of this approach is, instead of modifying the labels, it assigns different weights to the examples based upon their categories of protected attribute and outcome such that bias is removed from the training dataset. The weights are based on frequency counts. However as this technique is designed to work only with classifiers that can handle row-level weights, this may limit your modeling options" [8]. We used a logistic regression model for this algorithm, as it is the easiest to interpret in the given context. After running the fairness metrics using the pre-processing algorithm, we were able to compare our results to the baseline metrics from the previous section.

In-processing mitigates bias in classifiers while building a model. A classifier "is an algorithm that automatically orders or categorizes data into one or more sets" [**?**]. The in-processing technique we use is the prejudice remover algorithm, which accounts for the fairness metric as part of the input and returns a classifier optimized by that particular metric. In order to do this, we first needed to convert our data frame into a data type called a BinaryLabelDataset. Similarly to pre-processing, we compared the results of our in-processing methods with both the baseline and the pre-processing to gauge which method so far has better individual or group fairness.

The prejudice remover is a method for reducing indirect prejudice (i.e. how COMPAS is racially biased because it uses proxy variables for race). The prejudice remover implements two different regularizers, one to avoid overfitting and one to enforce fair classification. The prejudice remover regularizer works by minimizing the prejudice index, a mathematical equation for quantifying fairness defined by Kamishima et. al. This in turn enforces a classifier's independence from sensitive information (e.g., race).

Our last approach, post-processing bias mitigation, is used after training a model. Post-processing algorithms equalize the outcomes (i.e., predicted classification values) to mitigate bias instead of adjusting the classifier or the training data [10]. We use calibrated equalized odds, which "optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective" [insert aif360 website citation here]. An equalized odds objective constrains classification algorithms such that no error type (false-positive or false-negative) disproportionately affects any population subgroup; both groups, in our case both white and Black defendants, should have the same false-positive and false-negative rates. Through the calibrated equalized odds method,

we want to decrease bias while also maintaining calibration [11]. Calibration refers to improving a model so that the distribution of predicted outcomes is similar to the distribution of observed probability in the training data.

4. Results

With our baseline model, we ran the four different group fairness metrics we chose and compared the results (pictured in Table 2).

The statistical parity difference is -0.14. This indicates that there is a large difference between white and Black defendants regarding whether or not they recidivate. The algorithm unfairly benefits white defendants over Black defendants. Disparate impact ratio is 0.47. The ratio of white defendants predicted to not recidivate to the Black defendants predicted to not recidivate is 0.47. A ratio between 0.8 and 1.25 is considered fair, therefore the algorithm unfairly benefits white defendants.

Average odds difference is -0.44. The average difference in false positive rates and true positive rates for white and Black defendants is -0.44. Values less than zero are considered in favor of the privileged group, so the algorithm unfairly benefits white defendants.

Equal opportunity difference is -0.41. The difference of true positive rates between the Black and white groups is -0.41. A value less than 0 indicates a benefit to the privileged group, so the algorithm unfairly benefits white defendants. The value is substantially less than -0.1, which indicates that the algorithm benefits white defendants.

All four group fairness metrics determine that the COMPAS algorithm favors white defendants over Black defendants. Although the magnitudes of the various fairness metrics are different, all of the fairness metrics are not within their respective fairness thresholds. Our goal is to use pre-processing, in-processing, and post-processing algorithms in the AIF360 toolkit to see if we can make COMPAS fair at all.

Table 2. Results of Baseline Fairness Metric Analysis

Fairness Metric	Ideal Value	Baseline Value	Benefited Group
Statistical Parity Difference	0	-0.14	White Defendants
Disparate Impact Ratio	1	0.47	White Defendants
Average Odds Difference	0	-0.44	White Defendants
Equal Opportunity Difference	0	-0.41	White Defendants

4.1. Pre-Processing Approach

After running the reweighing algorithm, our fairness metrics are -0.015 for statistical parity difference, 0.015 for equal opportunity difference, 0.014 for average odds difference, and 0.98 for disparate impact ratio. Overall, these fairness metrics show that the pre-processing algorithm reweighing improves the bias in the COMPAS algorithm.

References

1. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
2. Dieterich, W.; Mendoza, C.; Brennan, T. Response to ProPublica: Demonstrating accuracy equity and predictive parity. *equivant* **2016**. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.
3. Larson, J.; Mattu, S.; Kirchner, L.; Angwin, J. How we analyzed the COMPAS recidivism algorithm, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
4. <AI & Equality>: A Human Rights toolbox. https://colab.research.google.com/drive/11SWijBDkr1pSgIXyjUkgoSB0gh5oAlaI.

5. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; Nagar, S.; Ramamurthy, K.N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K.R.; Zhang, Y. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, 2018. https://arxiv.org/abs/1810.01943.
6. Ashokan, A.; Haas, C. Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management* **2021**, *58*, 102646.
7. Kypraiou, S. What is Fairness?, 2021. https://feministai.pubpub.org/pub/what-is-fairness-/release/1.
8. Ronaghan, S. AI Fairness — Explanation of Disparate Impact Remover, 2019. https://towardsdatascience.com/ai-fairness-explanation-of-disparate-impact-remover-ce0da59451f1.
9. Cortez, V. How to define fairness to detect and prevent discriminatory outcomes in Machine Learning, 2019. https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2.
10. Baxter, K. What is AI bias mitigation, and how can it improve AI fairness? **2021**.
11. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; Weinberger, K.Q. On fairness and calibration. *Advances in neural information processing systems* **2017**, *30*.