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 4, 2022 submitted to Journal Not Specified

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. 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. published a rebuttal on behalf of Northpointe defending the COMPAS algorithm and refuting Angwin et al.'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. Our research found that the pre and post-processing bias mitigation algorithms, specifically reweighing and calibrated equalized odds, are the most effective at improving fairness.

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. 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" [2]. After examining the fairness metric statistical parity difference, Angwin et al. conclude that the algorithm is racially biased [2].

Dieterich *et al.* [3], on behalf of Northpointe, deny the allegations of racial bias and offer their own analyses based on different fairness metrics in rebuttal [3]. Angwin *et al.* [1] maintain that there are biases in the outcome values, protected attributes, and covariates during Dieterich *et al.* [3]'s data processing phase. ProPublica collaborators Larson *et al.* [2] account for these biases in their analyses. In their response, Dieterich *et al.* [3] 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 debiasing 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 debiasing algorithms used in its AI & Equality Human Rights Toolbox to conduct our own analyses on the COMPAS dataset. 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 debiasing 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 library addition is the COMPAS General Recidivism Risk Scores dataset from the AI Fairness 360 (AIF360) toolkit. The AIF360 toolkit builds on the dataset released by ProPublica collaborators Larson et al. created to examine the racial bias and the true outcomes of the recidivism risk scores in the COMPAS algorithm for the initial "Machine Bias" article. For this, Larson et al. obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida, as well as the corresponding intake information for each defendant including but not limited to name, sex, race, age, and charge degree and description. They also obtained data about whether defendants actually recidivated or not in the two year period following their initial COMPAS score assessment. AIF360 then processed the data with the same procedures that Larson et al. followed for their analysis.

The raw data that we use from the AIF360 has 6,167 rows, where each row represents an arrest charge for a defendant. AIF360's COMPAS data includes the defendant's age, race, sex, what they were charged with, and whether or not the defendant ultimately recidivated within a two-year period after their arrest.

This exploration aims to evaluate anti-Black algorithmic bias and the differing effects of the COMPAS algorithm between white and Black defendants; as such, we filter the data to only include

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 dataset we will be using, containing information on a defendant's age, sex, race, criminal history, charge degree, charge description, and two-year recidivism outcome.

individuals whose race is listed as Caucasian or African-American. Our data therefore has 5,723 rows with information on defendant race, age, gender, prior crimes, and two-year recidivism rate (Figure 1). The distributions of age (Figure 2), prior charges (Figure 3), and recidivism (Figure 4 and Table 1) all vary by race. The average African-American defendant in our dataset is 29 years old, male, and has committed two prior crimes. The average Caucasian defendant in our dataset is 35 years old, male, and has committed one prior crime. The average defendant for both races does not have any juvenile convictions.

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

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 datasets, including the COMPAS dataset 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 use to debias the COMPAS algorithm [5]. Fairness metrics are mathematical measures of whether an algorithm treats members of different groups (such as racial or gender groups) equally. An algorithm has no understanding of the historical oppression of certain groups and how such bias is baked into the data on which the algorithm is trained. Thus, fairness metrics provide a method of evaluating an algorithm's level of bias towards or against the unprivileged group versus the privileged group. We researched the definitions and applications of different fairness metrics [6] to determine which metric would be most appropriate for our project. To begin, we implemented a logistic regression model predicting recidivism using the defendant data. This analyzes the baseline values without any bias mitigation. We use logistic regression, as it is the easiest model to interpret in the given context.

Fairness is subjective; one person might consider an algorithm fair if groups are given the same treatment, while someone else might only consider the algorithm fair if groups as a whole receive the same outcomes. In our research we choose to use the definition of group fairness as our definition for assessing fairness in the processing approaches. 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].

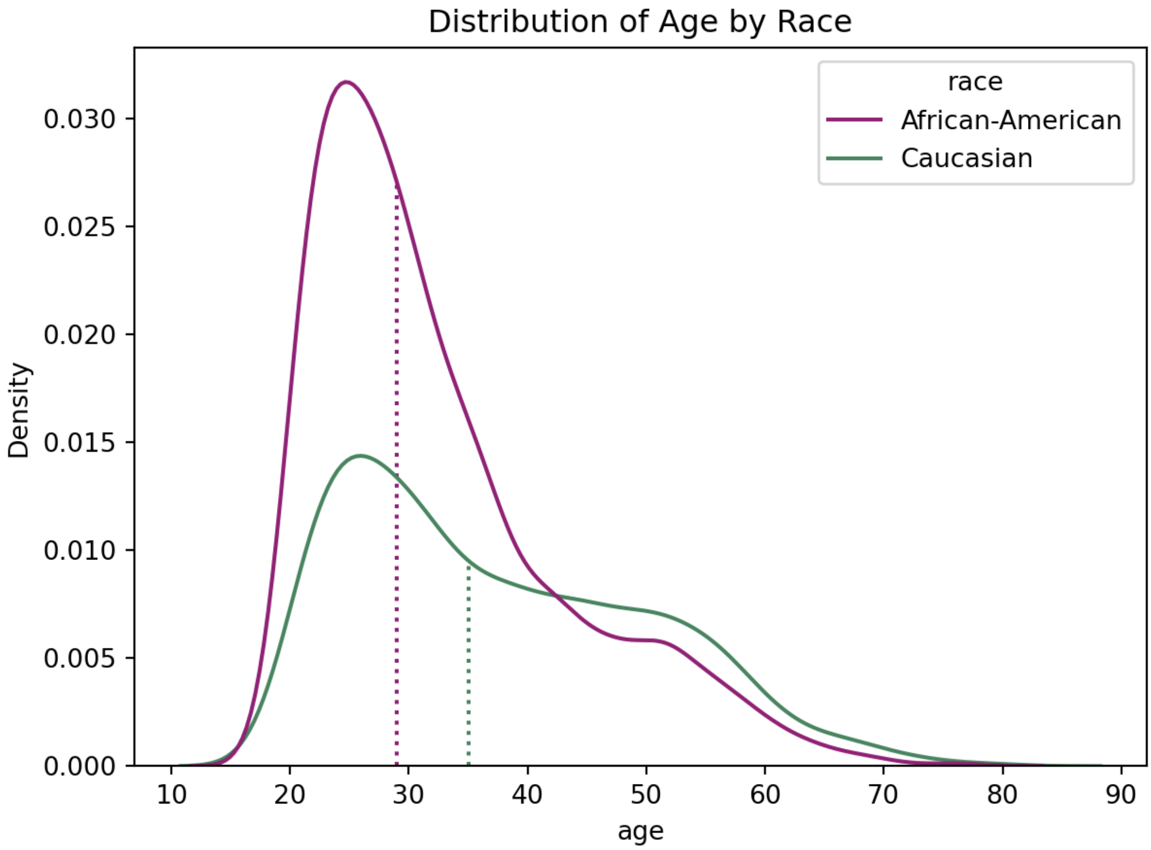

Figure 2. The purple curve shows the distribution of the ages of Black defendants, and the green curve shows the distribution of the ages of white defendants. The probability of a defendant's age being between two points on the x-axis is the total shaded area of the curve under the two points. The purple 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.

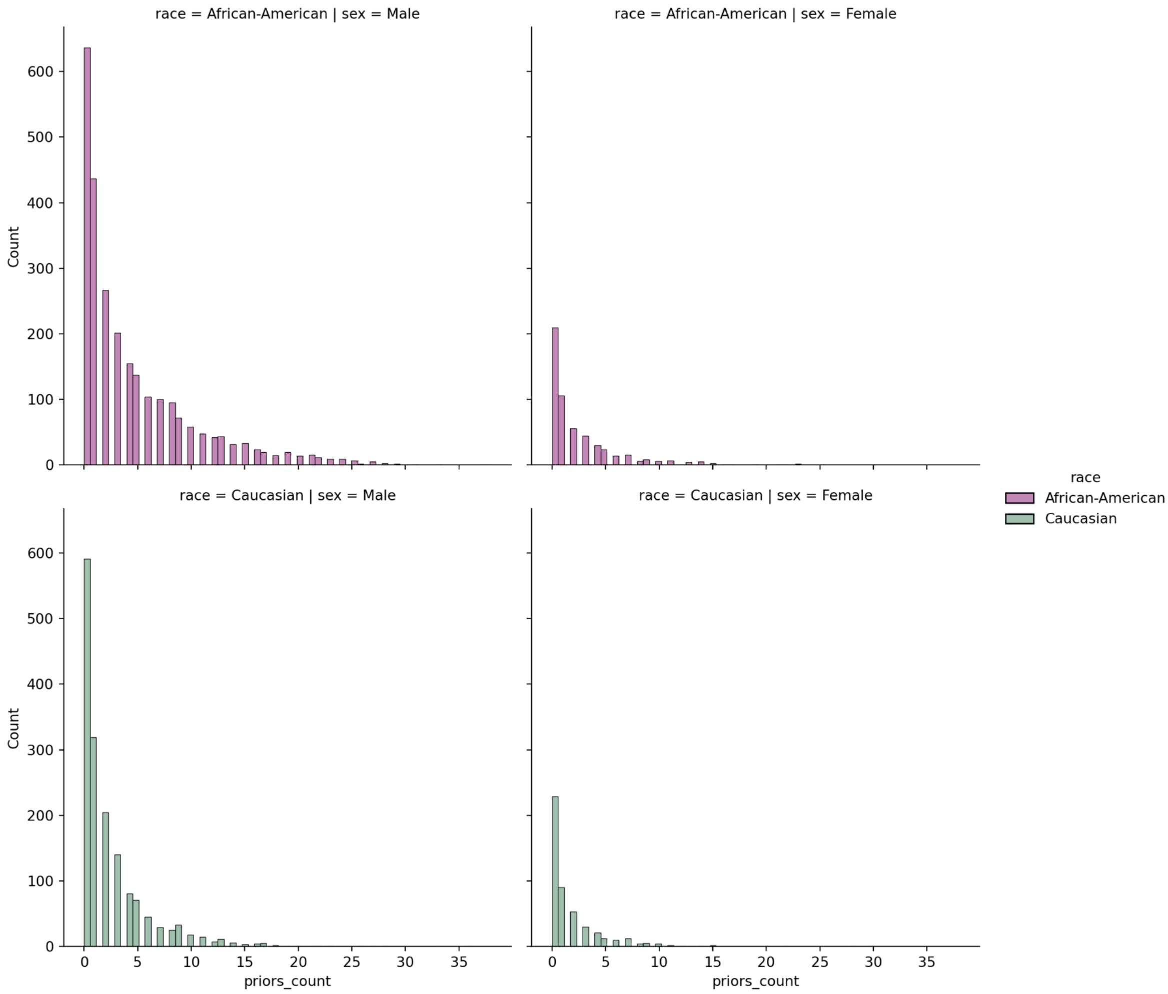

Figure 3. Black defendants, particularly men, are more likely to have a greater count of prior charges than white defendants. Male defendants have a higher number of prior charges than do female defendants. Though we do not know for sure which information goes into the COMPAS algorithm, it is likely 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.

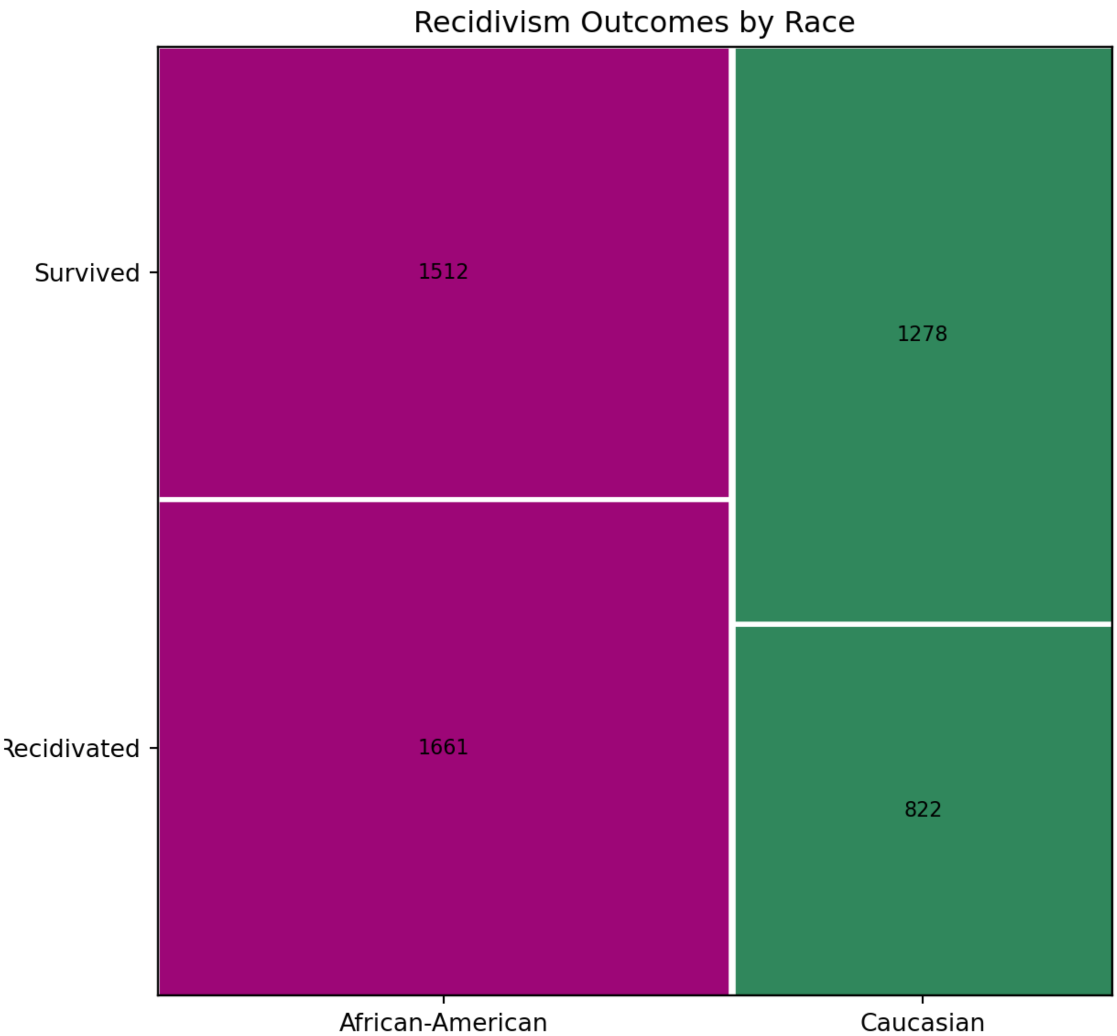

Figure 4. When we divide the data into Black and white defendants, we can see that Black defendants recidivate more than white defendants and Black defendants are more likely to recidivate than not recidivate. 39.14% of white defendants did recidivate within two years compared to 52.35% of Black defendants.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

Group and individual metrics work in opposition of one another, meaning that when group fairness improves, individual fairness gets worse [7]. We chose the following four group fairness metrics to evaluate our models.

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

The equal opportunity difference 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 dataset, 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 sought to determine where in the data science pipeline we can mitigate the most bias, using pre-processing, in-processing, and post-processing debiasing algorithms. These are all based on using predictive models to figure out how we can "fix" the bias that is present.

3.5. Pre-Processing

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]. 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.

3.6. In-Processing

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" [10]. 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.

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 [11]. The prejudice remover regularizer works by minimizing the prejudice index, a mathematical equation for quantifying fairness defined by Kamishima *et al.* [11]. This in turn enforces a classifier's independence from sensitive information (e.g., race). Similar to pre-processing, we compare the results of our in-processing methods with both the baseline and the pre-processing model to gauge which method so far has the better group fairness.

3.7. Post-Processing

Our last approach, post-processing bias mitigation, is implemented after training a model. Post-processing algorithms equalize the outcomes (i.e., predicted recidivism 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" [5]. 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 [12]. Through the calibrated equalized odds method, we want to decrease bias while also maintaining calibration [12]. 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 Figure 5).

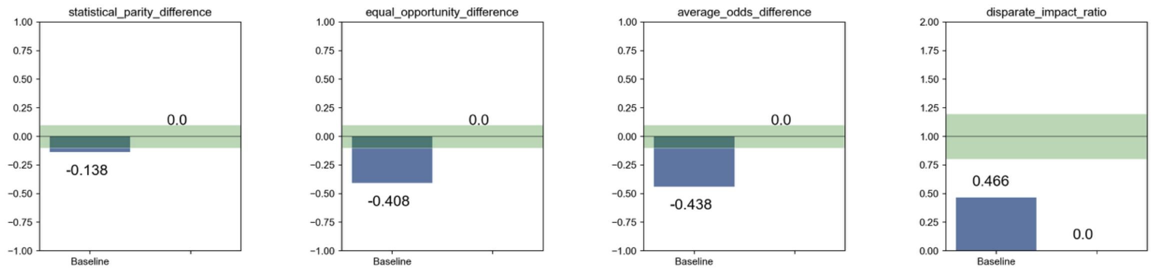

Figure 5. Fairness Metrics of Baseline Model. For each of the four fairness metrics, the green section represents the range of fair values. The blue bars indicate the values of each of the metrics for the baseline model.

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 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 benefits white defendants. The value is substantially less than -0.1, which indicates that the algorithm is unfair.

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, none of the metrics are 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.

Our confusion matrix for the baseline model (see Figure 6) indicates that the false positive and false negative percentages are 16.68% and 22.06%, which are the lowest values for this matrix. The highest value is the value of true positives, at 35.41%, showing that of the total number of predictions, 35.41% were correct predictions of recidivism.

4.1. Pre-Processing Approach: Reweighing

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. All of these values now fall within the margins of fairness (see Figure 7). These values make sense because the reweighing algorithm chooses different weights depending on whether an attribute is protected. This does an effective job of eliminating the bias, shown by each of the values in the graph being within the "fair" range of the particular fairness metric. Overall, these fairness metrics show that the reweighing algorithm improves the bias in the COMPAS algorithm.

The highest percentage in the reweighing confusion matrix (Figure 8) is for the percentage of true negatives, defendants who were predicted to not recidivate and actually did not, at 36.24%, 0.83% higher than the true negatives in the baseline model. This means that the reweighing model marginally improved the baseline model accuracy in predicting people who did not recidivate. However, the percentage of true positives, or those who were predicted to recidivate and did recidivate, is 24.72%, 1.13% lower than that of the baseline model. This means that the reweighing model slightly lowered the accuracy of the baseline model in predicting people who did recidivate. As a result, the overall

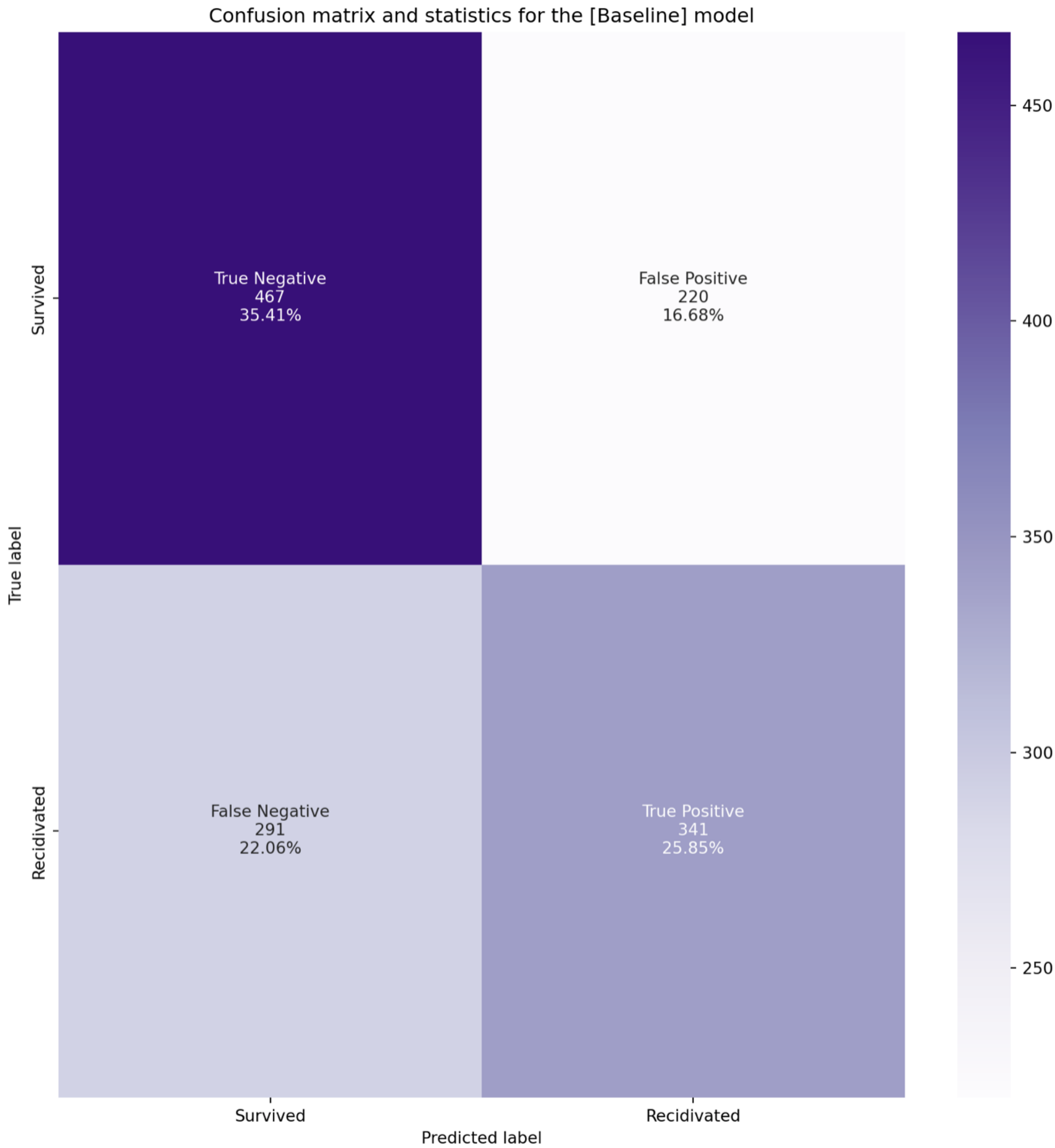

Figure 6. Confusion Matrix of the Baseline Model

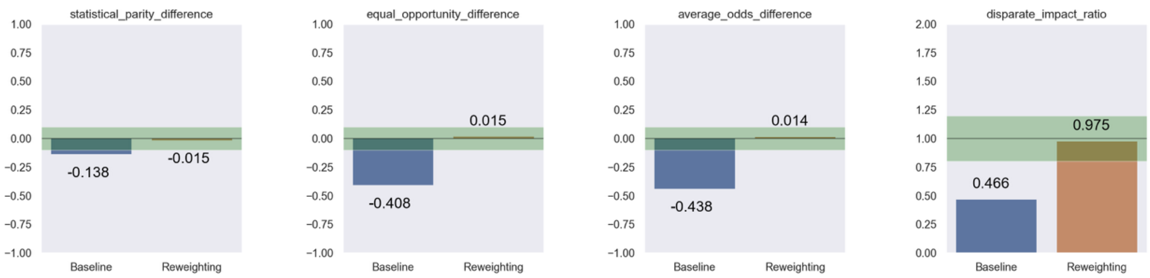

Figure 7. Fairness Metrics of Reweighing Model. For each of the four fairness metrics, the orange bars indicate the values for the reweighing model.

reweighing model accuracy remains about the same as the baseline model accuracy. The false negative and false positive percentages, or the predictions that proved to be incorrect, are 23.20% and 15.85%, respectively. The re-weighing model increased the number of false negatives by 1.14% and decreased the number of false positives by 0.83%. We are most concerned about the false positives, which indicate defendants who are predicted to recidivate and do not actually recidivate. While there is a slight improvement in the percent of false positives, it is still relatively high at 15.85%, indicating that we should further precede with additional bias mitigation techniques.

4.2. In-Processing Approach: Prejudice Remover

Our fairness metrics after running the prejudice remover are 0.367 for statistical parity difference, 0.321 for equal opportunity difference, 0.344 for average odds difference, and 3.3 for disparate impact ratio. Like the model performance metrics suggested, the prejudice remover approach resulted in an increased benefit for Black defendants. As the orange bars on Figure 9 show, the values of the fairness metrics have reversed from their baseline values. Now all four metrics suggest an unfair advantage for Black defendants. Thus, this approach removed the model's prejudice against Black people, but it did not result in a "fair" model.

The highest percentage in this confusion matrix (Figure 10) is for false positives, defendants who the model predicted to recidivate and actually did not, at 37.76%. This false positive percentage concerns us because we do not want defendants who do not recidivate to have unfairly long sentences due to their (incorrectly) predicted recidivism. The number of true positives, defendants who the model predicted to recidivate and actually recidivated, is lower at 23.88%. The percentage of false negatives is lower, whereas the true negative is low at 14.33%. This indicates that the model does not do a good job in predicting the defendants who don't recidivate. It is very likely that it will incorrectly predict that someone will recidivate.

After the prejudice remover, the accuracy of the model is fairly similar across the board of races, but the model's accuracy is only 38.21%, which is much lower than the baseline and reweighing models. This accuracy score means that the model makes accurate predictions only 38% of the time.

4.3. Post-Processing Approach: Calibrated Equalized Odds

After running the calibrated equalized odds algorithm, our fairness metrics are 0.134 for statistical parity difference, 0.088 for equal opportunity difference, 0.146 for average odds difference, and 1.155 for disparate impact ratio. As we can see in Figure 11, the calibrated equalized odds approach results in all four metrics now suggesting a benefit for the originally unprivileged group, Black defendants. Though the values for statistical parity difference and average odds difference are slightly above the range of what is considered fair, the margin is much smaller than the original margin between the value and the range of fairness (as indicated by the blue bars). Thus, the calibrated equalized odds approach successfully counteracts the bias against Black defendants and results in a mostly fair model.

The highest percentage in this confusion matrix (Figure 12) is for true negatives, defendants who the model predicted to not recidivate and actually did not, at 50.04%. The number of true positives, defendants who the model predicted to recidivate and actually recidivated, is much lower at 3.11%. The percentage of false negatives is somewhat high for this baseline model, whereas the amount of false positives is very low at 2.05%. This indicates that the model does a good job in predicting the defendants who don't recidivate. It is very unlikely that it will incorrectly predict that someone will recidivate. Though the accuracy score for the model is 0.53, lower than the baseline, this model is much better at identifying true negatives.

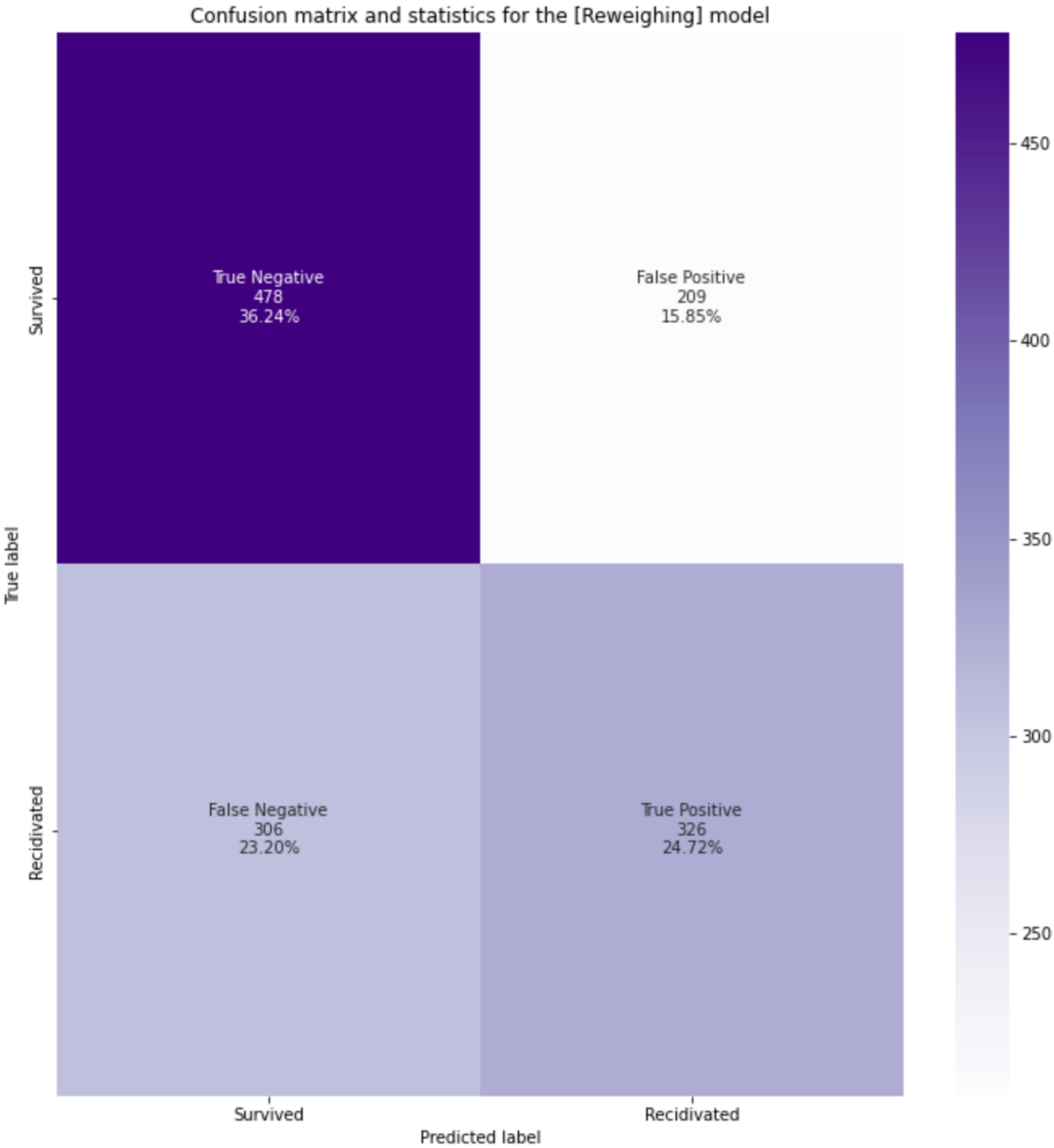

Figure 8. Confusion Matrix of the Reweighing Model

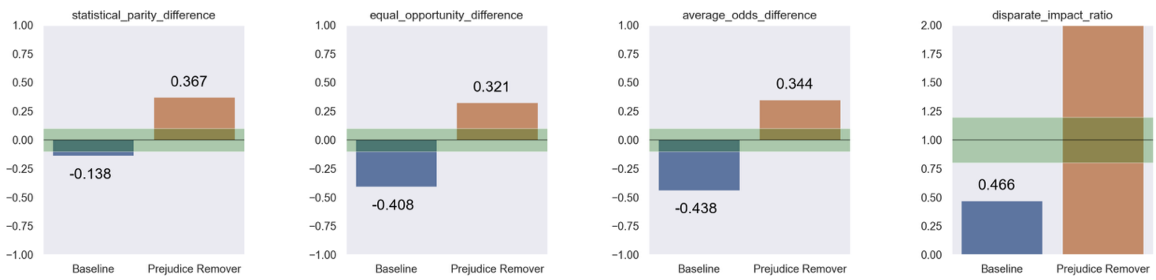

Figure 9. Fairness Metrics of Prejudice Remover Model. For each of the four fairness metrics, the orange bars indicate the values for the prejudice remover model.

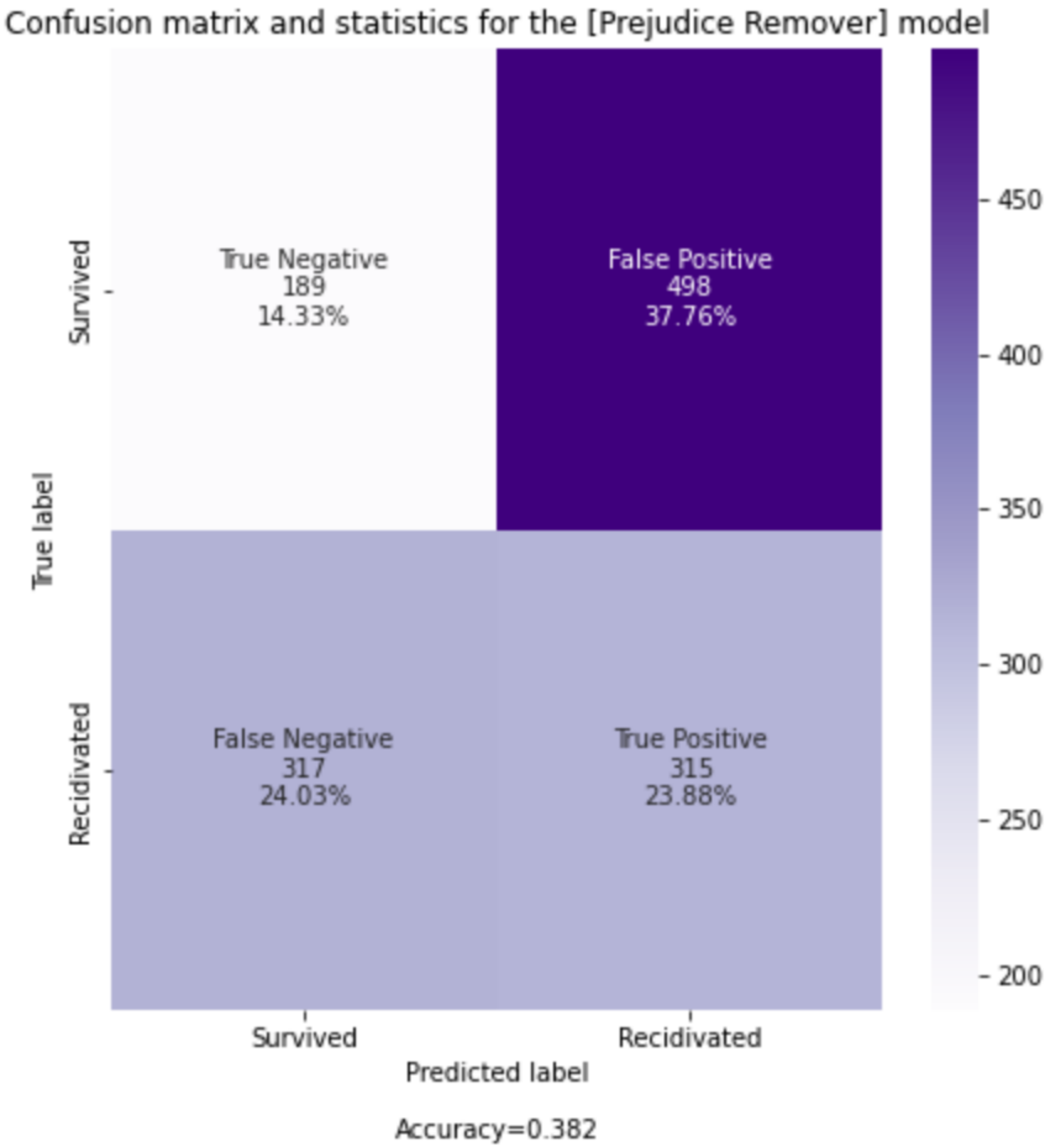

Figure 10. Confusion Matrix of the Prejudice Remover Model

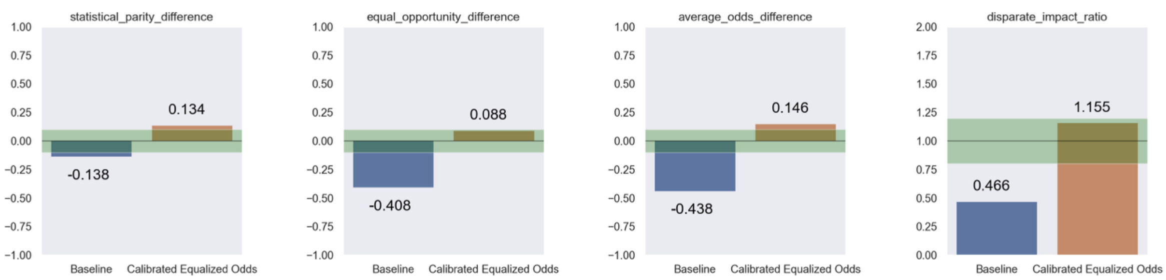

Figure 11. Fairness Metrics of Calibrated Equalized Odds Model. For each of the four fairness metrics, the orange bars indicate the values for the calibrated equalized odds model.

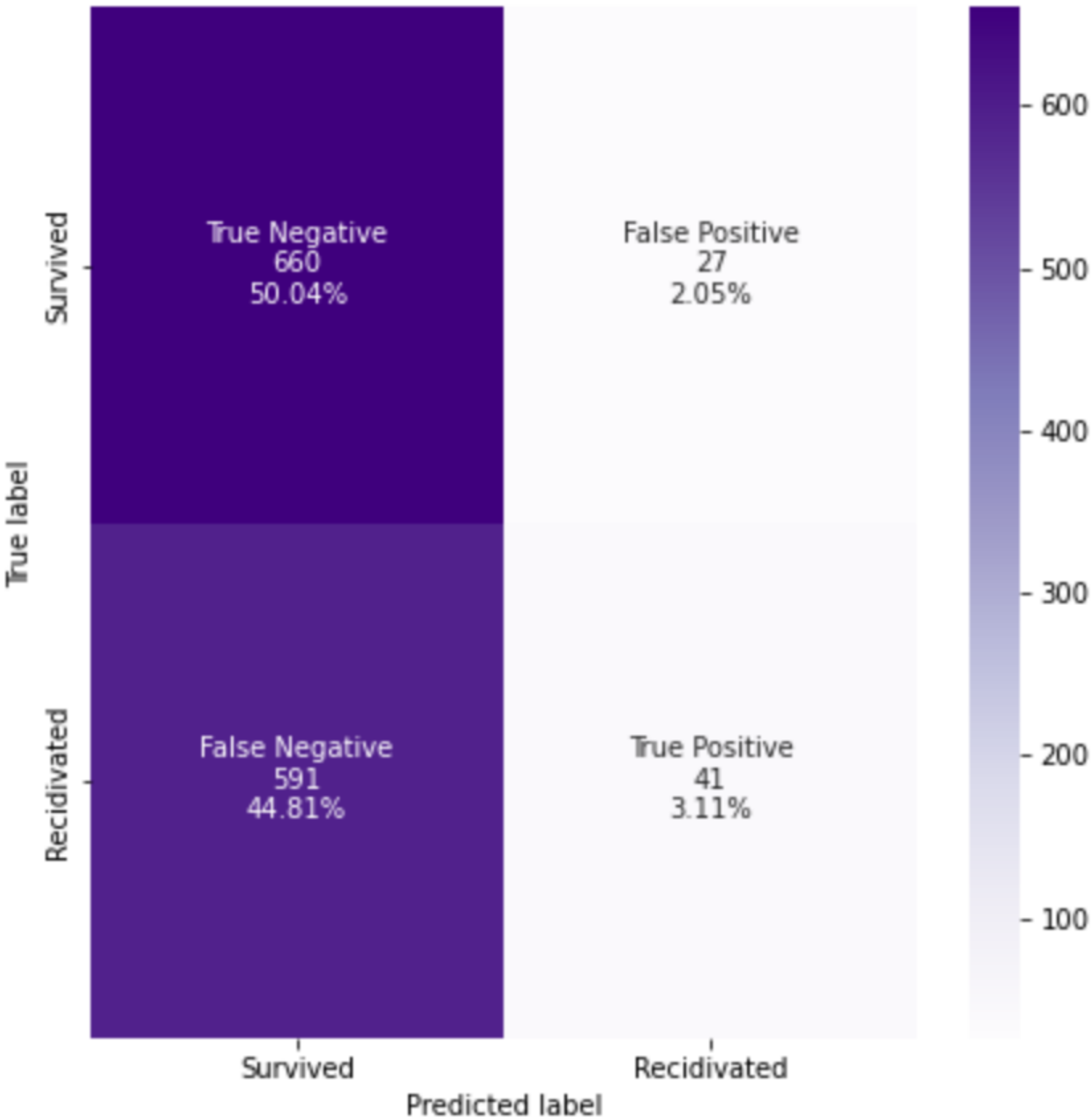

Figure 12. Confusion Matrix of the Calibrated Equalized Odds Model

5. Conclusion

5.1. Limitations

Our project tries to mitigate bias within existing data, rather than within the methods used to collect this data. The data are collected using a risk assessment survey, where defendants are asked a series of questions (see introduction) which are supposedly used to determine whether someone will recidivate or not. Many of these questions involve proxy variables for race. The creation and facilitation of this survey can therefore be assessed for bias mitigation as the methods used may potentially privilege some groups over others. As a result, the bias exists even before values are collected, limiting the debiasing work able to be done on the output values.

Furthermore, AIF360 processed the data we used in this project the same way ProPublica processed their data. AIF360 simplifies the original defendant data and their COMPAS scores to make the data easier to analyze. However, this initial processing may lose some details captured in the original raw data. As we have noted throughout this paper, bias can be introduced in any step of the data science pipeline and AIF360's initial data processing step is no exception.

AIF360 contains an array of processing methods that have been written by various scholars of fairness [5]. These methods are each distinct and apply a different technique to mitigate bias. Our results only contain three processing methods, one for each step in processing. This choice was based on time constraints as well as our working environment. For example, the code for most of the other in-processing methods would not run for us. In-processing on its own works to debias an algorithm, not the input or output values, so some of those methods may not work on certain algorithms. We had similar challenges with pre and post-processing, where the methods were challenging to run and would require more time to be spent debugging in order to make them work. Had we used alternative processing methods, our results would be quite different, as the method changes apply a different mathematical technique to the input.

5.2. Future Work

In this project, we have only focused on a few select fairness metrics and a few debiasing algorithms. We looked at group fairness metrics because we specifically wanted to mitigate systemic racial bias in the COMPAS algorithm, but other researchers could compare how well debiasing algorithms work using individual fairness metrics. Future work could also include conducting more tests using other debiasing algorithms to see if the COMPAS algorithm could be made more fair. One significant limitation of our work is that some of the in-processing algorithms provided in the AIF360 toolkit were difficult to integrate into our code and we were unable to run many of them. The processing methods we chose, though effective, were based on our work environments, and there are other existing algorithms that attempt to mitigate bias.

We only looked at defendants labeled as Caucasian and African-American in our dataset, which diminishes the generalizability and nuance of our results. Comparing how COMPAS assesses defendants of other races would be an incredibly relevant extension of our existing work. Furthermore, Larson *et al.* [2] also found that age was the most predictive factor of a higher risk score: "Defendants younger than 25 years old were 2.5 times as likely to get a higher score than middle aged offenders, even when controlling for prior crimes, future criminality, race and gender" [2]. In addition, female defendants were more likely to receive a higher recidivism risk score than male defendants when controlling for the aforementioned factors [2]. Thus, some extensions of our work could include attempting to remove age and gender bias and determining the most effective processing algorithms to do so.

The dataset we are working with only contains defendants' criminal history records from Broward County, Florida. Many other states, including New York and Massachusetts, use COMPAS to calculate recidivism scores, and a further study could apply this paper's methods to the data coming out of other states. By conducting experiments on multiple states the baseline model would be trained on a larger and more diverse set of data, which could improve its accuracy. In addition, by adding more geographically diverse data to our set we can start to create more generalizable conclusions about our results. However, future researchers must keep in mind the real-life consequences of recidivism risk algorithms like COMPAS and how research on methods of debiasing algorithms could potentially be utilized to justify the continued use of biased algorithms.

5.3. Final Thoughts

Through our methods, we concluded that the pre-processing and post-processing methods mitigate bias and increase fairness most successfully. The in-processing technique was effective at mitigating bias, but did not result in a mathematically fair model. However, these techniques only treat the symptoms of racial bias in the justice system, as opposed to addressing the root cause, systemic racism. We need to question why these flawed algorithms exist and why they can be used to determine someone's fate in the justice system. While we may have methods for reducing biased algorithms such as COMPAS, the focus should be on addressing the over-policing and over-criminalization of Black communities that result in biased data and algorithms.

5.4. Ethical Statement

Throughout this project, we endeavored to debias an incredibly powerful algorithm that can change the course of someone's life. However, we want to recognize that debiasing methods are not the most effective or beneficial way to uproot bias within the American justice system. The problems within the justice system are much more complex than an algorithm and are rooted in the United States's history of racism. In addition, algorithmic bias is much more complex than the algorithm itself; algorithmic bias comes from the people who produce the algorithm. We all have our own beliefs, biases, and situated knowledge that impact and subsequently limit everything we create.

6. Acknowledgements

We would like to thank our sponsor organization Women At The Table and Professor Ben Baumer for introducing us to this incredibly meaningful work, supporting us, and answering the many questions we had throughout this semester-long project. We would specifically like to thank Sofia Kypraiou for her mentorship and patience throughout this process.

Method	koRpus	stringi
Word count	4719	4537
Character count	29868	29825
Sentence count	280	Not available
Reading time	23.6 minutes	22.7 minutes

References

1. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine bias. Report, ProPublica, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
2. Larson, J.; Mattu, S.; Kirchner, L.; Angwin, J. How we analyzed the COMPAS recidivism algorithm. Report 1, ProPublica, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
3. Dieterich, W.; Mendoza, C.; Brennan, T. Response to ProPublica: Demonstrating accuracy equity and predictive parity. *Northpointe Inc. Research Department* **2016**. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

4. Kypraiou, S. <AI & Equality>: A Human Rights toolbox. Resource, Women @ the Table. 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. Resource, AI Fairness 360, 2018. https://doi.org/10.48550/arXiv.1810.01943.
6. Ashokan, A.; Haas, C. Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management* **2021**, *58*, 102646. https://doi.org/10.1016/j.ipm.2021.102646.
7. Kypraiou, S. What is Fairness? *From Bias to Feminist AI* **2021**. https://feministai.pubpub.org/pub/what-is-fairness-/release/1.
8. Ronaghan, S. AI Fairness — Explanation of Disparate Impact Remover. Editorial, Towards Data Science, 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. Editorial, Towards Data Science, 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? Editorial, InfoWorld, 2021. https://www.infoworld.com/article/3630450/what-is-ai-bias-mitigation-and-how-can-it-improve-ai-fairness.html.
11. Kamishima, T.; Akaho, S.; Asoh, H.; Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*; Springer, 2012; pp. 35–50. https://doi.org/10.1007/978-3-642-33486-3_3.
12. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; Weinberger, K.Q. On fairness and calibration. *Advances in neural information processing systems* **2017**, *30*. https://doi.org/10.48550/arXiv.1709.02012.