# ANALYZING THE COMPAS ALGORITHM IN CRIMINAL DEFENDANT RISK ASSESSMENT

YASMINE AYAD
ADVISOR: SORELLE FRIEDLER

Contents



## 1. Abstract

For my thesis, I analyzed the COMPAS recidivism prediction tool made by Equivant which aims to see how likely a defendant charged with a crime will re-offending given a score from 1-10 where 1 indicating lowest risk and 10 indicating highest risk and is used by many states in the country. ProPublica dataset consisted of re-arrest data of COMPAS predictions of 6172 people made between 2012-2014 which they proved that COMPAS was more likely to falsely label African-American defendants as high risk more often than White defendants and more likely to falsely label White defendants as low risk more often than African-American defendants. Looking at ProPublcia's dataset along with Jai Nimgaonkar's dataset who took ProPublica's dataset to see if these people were convicted of a crime in order to see if there is still bias from re-arrest data to conviction data when looking at intersectionality between sex and race and different fairness aware algorithms.

## 2. Introduction

There are many situations and settings where a decision is to be made taking in some observable set of features and weighing in the risks of each situation to make a final decision. These judgments are now taking in a risk assessment derived from an algorithm and statistical frameworks in order to make a decision on what the next steps should be. Risk assessment is a method to identify hazards and risk factors, analyze and evaluate the risk associated with that hazard and determine appropriate ways to eliminate or control the hazard or make appropriate decisions with the analysis. Risk assessment tools and algorithms being used in different fields have been controversial in that there has been criticism that these tools are being biased and discrimination as they are trained on large datasets of past instances. Let us look at three examples where this may arise in different fields. One example is in some parts of the criminal justice system in the United States, risk assessment tools are used to assess a defendant's probability of recidivism or future arrests based on their past history and other attributes. The result of this tool is then used to make decisions on bail, sentencing or parole, more on how this achieved and how it works discussed later. Another example being analyzed using these such decision tools is the different ways in which genders and racial groups experience advertising and commercial content on the Internet differently. That is, for example, if a male or female user is equally interested in a particular product, does it follow that they're equally likely to be shown an ad for it? This could lead to bias in that women are shown advertisements for lower-paying jobs or if the female is interested in football is the advertisement being shown to her the same as a male user, if at all. Harvard professor Latanya Sweeney [12] says in her paper that names that are typically associated with black people produce ads related to arrest records in Google's ads. Another study by Datta et al. [3] done in 2015 showed that when a user is female Google's ads will result in having fewer instances of an ad related to high paying jobs than when a user is male. The third field that risk assessment tools are used in is medical testing and diagnosis. Doctors making decisions about a patient's treatment may rely on tests providing probability estimates for different diseases and conditions. Bias here can result in the patients' past, race and gender can result in risk of the disease being high for the wrong disease as medical tests may play differently for different conditions that vary widely in frequency between these groups as Kleinberg et al. tell in their paper [8].

As stated above, risk decision or discrimination-aware algorithms and tools have been used in the criminal justice system in the United States which states a defendant's risk score on whether they will recidivate in the future depending on past events. The criticism here is that there has been evidence of bias of scores outputted by the algorithm between black and white defendants. ProPublica [7] released an article stating that the risk assessment tool COMPAS was being biased for risk of recidivism score between white and black defendants. The COMPAS algorithm, found in the article, that there are higher false positive rates and lower false negative rates for black defendants than for white defendants. ProPublica showed that the algorithm was being biased in that black defendants

were more likely to have an incorrect high score and white defendants were more likely to have an incorrect low score. ProPublica [6] first looked at more than 10,000 criminal defendants from Broward County, Florida and looked at their predicted recidivism rates with their actual rates over a two-year period. ProPublica chose the COMPAS algorithm because it was one of the most popular risk assessment tool for pretrial and sentencing in the US. They chose to use Broward County data because it's a large jurisdiction that uses the COMPAS algorithm in pretrial and release decisions and Florida has good open-records laws. They discarded scores that were assessed at parole, probation or other stages in the criminal justice system because Broward County primarily used COMPAS scores to determine to release or detain a defendant in pre-trial. COMPAS predicts the recidivism rate by having the defendants respond to a questionnaire that will be fed to COMPAS and generates scores of "Risk of Recidivism" and "Risk of Violent Recidivism". Scores 1 to 4 were labeled by COMPAS as "Low"; 5 to 7 as "Medium"; and 8 to 10 as "High". ProPublica defined recidivism as a new arrest within two years that does not count traffic tickets, municipal ordinance violations and people who were arrested for failing to appear at their court hearing as recidivism. For violent recidivism, they used FBI's definition of violent crime thus murder, manslaughter, forcible rape, robbery, and aggravated assault.

ProPublica found that the algorithm predicted recidivism correctly 61% of the time, but was only correct in its predictions of violent recidivism 20% of the time. They also found that the algorithm correctly predicted recidivism for black and white defendants at about the same rate, but misclassified the white and black defendants when examined over a two-year follow-up period. Along with the mentioned findings, ProPublica also showed that controlling for prior crimes, future recidivism, age and gender, black defendants were 45% more likely to be assigned a higher risk score than white defendants. Black defendants were twice as likely as white defendants to be misclassified as having a higher risk of violent recidivism, while white defendants were 63% more likely to have been misclassified to have a low risk of violent recidivism. Even controlling for prior crimes, future recidivism, age, and gender, black defendants were 77% more likely to have a higher risk violent recidivism score than white defendants. It has been 2 years since the ProPublica study, thus this research will have a basis of what ProPublica did and analyze 2 more years of data from Boward County records with the goal of analyzing the same data under a different measure - conviction instead of rearrest. I will be evaluating different papers in order to see what a classifier means to be biased in a risk assessment tool and what different approaches that have been used to achieve this goal. I then will be analyzing the new data generated for convictions within the two years. I will be analyzing intersectionality between race and sex and seeing if using conviction data will result in less bias rather than re-arrests. I will then be analyzing different classifiers talked about in the literature review to see if there is less discrimination with our convictions dataset compared to ProPublica's dataset and will these classifiers mitigate the bias.

2.1. **Motivation.** The motivation here is to analyze the risk assessment in the criminal justice system in order to analyze a different dataset in order to make COMPAS less biased between white and black defendants, especially to have equal false positive and false negative rates between white and black defendants. COMPAS has been used in many courts and parole decisions in order to see if the defendant has a low, medium or high risk of recidivism after being released. Since ProPublica has stated evidence that there is machine bias in COMPAS' algorithm and black defendants are statistically receiving higher risk scores even for smaller crimes compared to white defendants who did a bigger crime but received a lower score, the algorithm's assessment affects if black defendants do get bail, parole or set to jail thus it is affecting their lives. This thesis will be focused on analyzing data collected from the Florida State corrections database two years after the ProPublica study has been conducted. This data consists of the defendants from the original ProPublica study and seeing if they have been convicted of a crime or not. Unfortunately we do not have all the data of if the defendant was convicted of the original COMPAS screening crime. I will then analyze how different algorithms, specially Zafar et al.'s [9] and Calders and Verwer's [1] algorithms, compare when used with our data. These algorithms have already been researched to include the least bias as possible, thus we will be comparing each algorithm while seeing how well each of them do and the shortfalls of them. With more data throughout a longer time being analyzed, we will also be researching if ProPublica's original analysis is correct and see if there are any flaws in how they interpreted the data while compensating for their lack of knowledge.

## 3. Review of Literature

In order to analyze the data taken from the Florida State corrections database to see what types of risk assessment or discrimination-aware algorithms will yield less biased results, we first analyze how different researchers define the components of such classifiers. Once we have this, we analyze what it means for a discrimination-aware classifier to be biased and analyze all the errors that would arise when an algorithm is biased. Then we look at specifically what went wrong with COMPAS. Authors Kleinberg et al [8] and Chouldechova [2] talk about and analyze these terms in their respective papers. Once these definitions have been established, different algorithms and classifiers that have been written and researched in order to have a risk score with the least amount of bias and discrimination even though it sometimes it isn't feasible. These algorithms will then be compared and used with our data that was collected in order to see which algorithm works best if any, and any improvements there could be in each algorithm.

3.1. **Kleinberg.** Understanding what it means for a probabilistic classification to be fair to different groups is the first step to seeing why COMPAS is a biased algorithm between white and black defendants. Kleinberg et al.'s [8] study discuss what it means for a risk assessment tool to be biased. They formalize three fairness conditions that should be simultaneously obtained in order for it to be unbiased. However, they hypothesize that all three conditions cannot be satisfied simultaneously which is what COMPAS does not

do. They discuss that for a risk assessment tool to be unbiased or fair, it needs to be well calibrated within each group, balance for the negative class and balance for the positive class. A risk assignment algorithm is well calibrated if:

$$P(Y = 1|S, R = 0) = P(Y = 1|S, R = 1)$$

i.e for each group $R \in \{0, 1\}$ where 0 is the unprivileged group and 1 is the privileged group, and each class $Y \in \{0, 1\}$, where 0 is the negative class and 1 is the positive class, with associated score S, the expected number of people from the each group that belongs the positive class be equal. A risk assignment algorithm has a balance for the negative class if:

$$P(S|Y = 0, R = 0) = P(S|Y = 0, R = 1)$$

i.e. the average score assigned to people of group 0 who belong to the negative class should be the same as the average score assigned to people of group 1 who belong to the negative class. A risk assignment algorithm has a balance for the positive class if:

$$P(S|Y = 1, R = 0) = P(S|Y = 1, R = 1)$$

i.e. the average score assigned to people of group 0 who belong to the positive class should be the same as the average score assigned to people of group 1 who belong to the positive class.

These criteria must all be achieved together in order for the risk assessment tool to be unbiased and will be the general assumption in the papers I study and talk about later. Kleinberg et al. show that all three conditions can be achieved simultaneously when there are perfect prediction and equal base rates. In order to achieve perfect prediction suppose that for each feature vector $\sigma$, there are either $p_\sigma = 0$ or $p_\sigma = 1$ where $p_\sigma$ denotes the fraction of people with feature vector $\sigma$ who belong to the positive class. This means that each person's class label (positive or negative) is known for certain. In this case, all feature vectors $\sigma$ with $p_\sigma = 0$ to group $R$ with score $S = 0$ are assigned, and all $\sigma$ with $p_\sigma = 1$ to a group with score equal to 1.

In order to satisfy equal base rates suppose that the two groups studied have the same fraction of members in the positive class; that is, the average value of $p_\sigma$ is the same for the members of each group. This is then the base rate of the group with respect to the classification problem. In this case, a single group $R$ with score equal to this average value of $p_\sigma$ is created, and everyone is assigned to $R$. Kleinberg et al's study results show that except in highly constrained special cases such as when there are perfect prediction and equal base rates, it's not possible to satisfy the three constraints simultaneously if the fraction of users in the positive class differs between the two groups studied.

In the papers that will be mentioned, it is assumed that these fairness properties are met to assess the fairness of their algorithms. There is variable that will be in the set of groups, $R$, and in this case, there are two groups not privileged and privileged, thus $R \in \{0, 1\}$. There is an outcome indicator, $Y \in \{0,\ 1\}$ where 1 indicates that the individual will recidivate and 0 will denote they do not.

3.2. **Chouldechova.** Chouldechova's [2] paper is similar to Kleinberg's study in that Chouldechova agrees that the factors together contribute to an algorithm that is not biased. However, Chouldechova hones in on the COMPAS algorithm to see what exactly is wrong with it. Since this thesis is exploring and analyzing ProPublica's data and analysis of COMPAS, analyzing Chouldechova's paper will allow us to see what questions she asked to understand the fairness of COMPAS and incorporate those questions into our analysis. Chouldechova also uses the term high-risk score threshold, $s_{\mathrm{HR}}$, in her analysis of bias where defendants whose risk score is higher than the threshold will be referred as high-risk, while the remaining defendants will be considered low risk. The term predictive parity is defined as at a threshold what is the likelihood of recidivism of high-risk offenders is the same in each group for a score S = S(x). i.e.

$$P(Y = 1|S > s_{\mathrm{HR}}, R = 0) = P(Y = 1|S > s_{\mathrm{HR}}, R = 1)$$

Predictive parity at a given threshold $s_{\mathrm{HR}}$ requires that the positive predictive value (PPV) of the classifier $\widehat{Y} = 1_{S>s_{\mathrm{HR}}}$ with a score above the threshold to be the same between the unprivileged and privileged groups. She warns that calibration, which is the same definition as Kleinberg et al. for calibration, i.e.

$$P(Y = 1|S = s, R = 0) = P(Y = 1|S = s, R = 1)$$

and predictive parity are not the same since well-calibrated scores can fail to satisfy predictive parity at a given threshold.

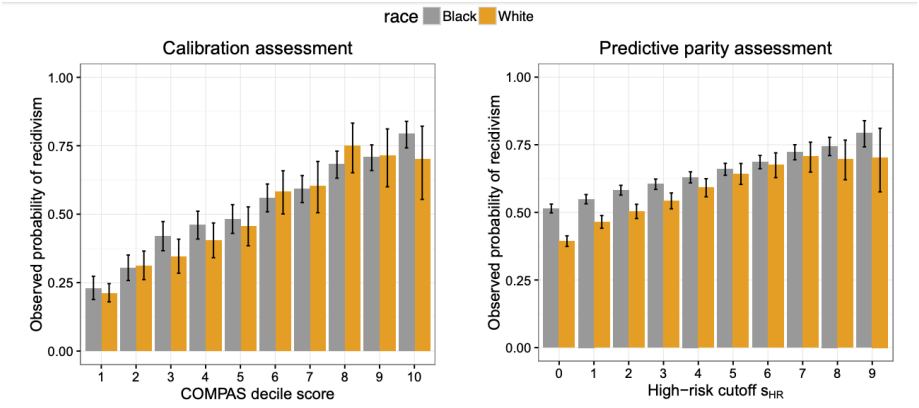
Figure 1 shows plots of the observed recidivism rates and error rates corresponding to the fairness notions of calibration, predictive parity and error rate balance, i.e.

$$P(S > s_{HR}|Y = 0, R = 0) = P(S > s_{HR}|Y = 0, R = 1)$$

and

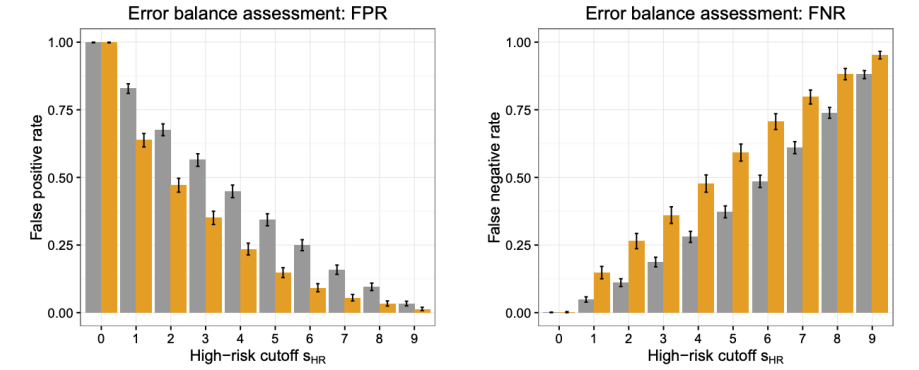
$$P(S \geq s_{HR}|Y = 1, R = 0) = P(S \geq s_{HR}|Y = 1, R = 1)$$

where the expressions in the first line are the group-specific false positive rates, and those in the second line are the group-specific false negative rates. These equations of false positive rates and false negative rates are similar to Kleinberg et al.'s balance for negative and positive classes with Chouldechova's equations for these error rates are determined at a given threshold $s_{HR}$. It is stated that COMPAS is well-calibrated and satisfies predictive parity when the high-risk threshold is 4 or higher. However, COMPAS fails on false positive and false negative error rate balance in the high-risk threshold as shown in Figure 1. The paper also shows that when recidivism prevalence is different between the two groups studied, the risk assessment tool that satisfies predictive parity at a given threshold must




(a) Bars represent empirical estimates of the expressions in (2.1): $\mathbb{P}(Y = 1 \mid S = s, R = r)$ for decile scores $s \in \{1, \ldots, 10\}$.

(b) Bars represent empirical estimates of the expressions in (2.2): $\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$




(c) Bars represent observed false positive rates, which are empirical estimates of the expressions in (2.3): $\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$

(d) Bars represent observed false negative rates, which are empirical estimates of the expressions in (2.4): $\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$

FIGURE 1. Empirical assessment of the COMPAS RPI according to three of the fairness criteria from Chouldechova's results [2]. Error bars represent 95% confidence intervals. These figures confirm that COMPAS is well-calibrated, satisfies predictive parity for high-risk cutoff values of 4 or higher, but fails to have error rate balance.

have false positive or false negative error rates that are not balanced. Thus, predictive parity and error rates are mutually exclusive when recidivism prevalence is different across groups.

These findings relate to Kleinberg et al. in that they show that for a risk assessment tool to be well calibrated and have balance for the positive and negative class there needs to be equal base rates across the groups. Thus, both Kleinberg et al. and Chouldechova show that recidivism prevalence needs to be the same for both groups to have equal error

| | Low-Risk | High-Risk |
|---|---|---|
| $Y = 0$ | TN | FP |
| $Y = 1$ | FN | TP |

TABLE 1. T/F denote True/False and N/P denote Negative/Positive. Thus, FP is the number of false positives: individuals who are classified as high-risk but who do not re-offend. TN is the number of true negatives: individuals who are classified as low-risk and who do re-offend. FN is the number of false negatives: individuals who are classified as low-risk and do re-offend. TP is the number of true positives: individuals who are classified as high-risk and do re-offend.

rates and balance for the positive and negative classes. At this point then predictive parity at a given threshold will not be satisfied. As Kleinberg et al. stated, it is very hard for a classifier to be well calibrated and have equal positive and negative rates, except when there is an equal base rate, which in Figure 1 Chouldechova was able to show that COMPAS does not satisfy all of these conditions. Given a particular choice of $s_{HR}$, a risk assessment algorithm's or tool's performance is evaluated in terms of a confusion matrix, as shown in Table 1. All of the fairness metrics presented above can be thought of as imposing constraints on the values in the confusion matrix. Another constraint, which cannot be controlled directly, is imposed by the recidivism prevalence ($p$) within groups. The prevalence ($p$), positive predictive value (PPV) and false positive and negative error rates (FPR, FNR) are related such as:

$$FPR = (p/1-p)(1-PPV/PPV)(1-FNR)$$

This equation shows that an instrument satisfies predictive parity if the PPV is the same across groups, but the prevalence differs between groups, the risk algorithm or tool cannot achieve equal false positive and false negative rates across the groups.

Chouldechova then shows how the differences in false positive and negative rates can result in disparate impact. She specifically addresses use cases where high-risk individuals receive stricter penalties and thus considers that a defendant receieves a penalty of $\mathrm{t_{min}} \leq T \leq \mathrm{t_{max}}$ where $T$ is the penalty the defendant receives in terms of time between defined minimum and maximum amount of time a defendant can receive. A simple risk-based approach would assign penalties as follows:

$$\mathrm{T_{MinMax}}(s) = \begin{cases} \mathrm{t_{min}} & if s > \mathrm{s_{HR}} \\ \mathrm{t_{max}} & if s < \mathrm{s_{HR}} \end{cases}$$

She shows the extent of disparate impact quantitatively as:

$$\Delta = \Delta(y_1, y_2) \equiv \mathbb{E}(T|R = 0, Y = y_1) - \mathbb{E}(T|R = 1, Y = y_2)$$

where $\Delta$ means the measure of disparate impact. The equation shows the expected difference in sentence duration between defendants in different groups and potentially different outcomes $y_1$ and $y_2$ which are in the set $\{0, 1\}$. She also discusses two different Corollaries of this result, one for non-recidivists and one for recidivists. The first corollary is for non-recidivists and says among individuals who do not recidivate, the difference in the average penalty under the MinMax policy is

$$\Delta = (\mathrm{t}_{\max} - \mathrm{t}_{\min})(FPR_0 - FPR_1)$$

where FPR is the false positive rate either for unprivileged defendants or privileged defendants. The second corollary is for recidivists and says among individuals who recidivate, the difference in the average penalty under the MinMax policy is

$$\Delta = (\mathrm{t}_{\max} - \mathrm{t}_{\min})(FNR_1 - FNR_0)$$

where FNR means the false negative rate for unprivileged or privileged defendants.

With risk assessment tools that satisfy predictive parity where recidivism prevalence differs across the groups, it will generally be the case that the group with the higher recidivism prevalence will have a higher false positive rate (FPR) and lower false negative rate (FNR). From the equations of the corollaries, this would on average result in greater penalties for defendants in the group that has a higher prevalence for both recidivists and non-recidivists. When $t_{min} = 0$, there may be sanctions as an alternative for incarceration. Taking it further to when $t_{min} = 1$ she defines a probability that a defendant receives a sentence over some period of incarceration as $\mathbb{E}T = \mathbb{P}(T \neq 0)$. From these observations, Chouldechova states that a non-recidivist in the unprivileged group are $FPR_0/FPR_1$ time more likely to be imprisoned compared to a non-recidivists in the privileged group.

**Theorem 3.1.** *There are three ways to control the classifiers used in a risk assessment tool in order to have less discrimination or bias:*

*(i) Allow unequal false negative rates to retain equal PPV's and achieve equal false positive rates*

*(ii) Allow unequal false positive rates equal PPV's and achieve equal false negative rates*

*(iii) Allow unequal PPV's to achieve equal positive and false negative rates*

From previous findings, FPR is a linear function of FNR under constraints on PPV. A consequence of (i) would be that if PPV is fixed then there will be an increase in FNR in order to balance FPR. From Chouldechova's finding in order to get $FPR_0$ to match $FPR_1$, the value for $FNR_0$ would need to increase which would be a big drop in accuracy and strategy (ii) would also cause undesirable consequences in order to achieve no bias. Even though strategies (i) and (ii) reduce disparate impact for one subgroup, they may increase disparate impact for another subgroup. Chouldechova then argues that the best strategy would be (iii) as it results in a score that does not satisfy predictive parity but by

allowing the high-risk threshold to differ across groups predictive parity can be satisfied. Chouldechova's and Kleinberg et al.'s results show that in order to for a risk assessment tool to have less discrimination there need to be equal positive and negative rates especially when there are equal base rates. As seen Chouldechova shows how disparate impact can result from using a recidivism prediction tool that is known to satisfy predictive parity. Chouldechova addresses, however, that a limitation in the analysis is that there can be potential biases in the observed data that effect to draw valid inferences concerning the fairness of a risk analysis tool. In the paper the assumption of the outcome of the tool $Y$ is a suitable outcome measurement to asses fairness in the risk assessment tool, however the true outcome of interest in ProPublica's study is reoffense which is not what she observed. Thus there maybe be a fraction of defendants in her data where $Y = 0$ did re-offend. If this is the case and there are group differences in the rates at which offenders are caught, the findings of empirical fairness assessments may be misleading. A future direction she considers in understanding how such forms of data bias affect the ability to assess tools with respect to different fairness criteria.

3.3. **Calders and Verwer.** Kleinberg and Chouldechova discuss what it means for a recidivism assessment tool to be unbiased and how such a tool can result in discrimination and errors between different classes. Now we will be looking at the earliest classifiers from 2010 that checks if the dataset has discrimination or not. Understanding already observed classifier algorithms will allow us to see which are the most efficient and has the least errors in data of COMPAS based on recidivism. The classifiers in this section will allow us to see if our modified dataset has discrimination compared to the dataset used in ProPublica when the error rates are equal and how they have evolved. Looking at Calders and Verwer's [1] paper allows us to see what types of classifiers for fairness there were before COMPAS and how these classifiers have evolved throughout time with Kleinberg, Chouldechova and Zafar[9], talked about in the next section.

They explore three different kinds of classifiers based on the Naïve Bayes classifier. They use the notion that in discrimination-aware tools every group should have the same probability of being in the positive class calling this independency constraints. The paper has a few assumptions in that there is a labeled dataset D, binary class attribute $Y$ which takes values $\{0, 1\}$ and on binary sensitive attribute R with values $\{0, 1\}$ and the sensitive attribute has an unwanted correlation with the class attribute. They want to optimize predictions that are non-discriminatory and accurate with these assumptions. Discrimination in their paper is measured through the difference $P(Y = 1|R = 1) - P(Y = 1|R = 0)$ which is the discrimination score. Throughout their research they questioned if the sensitive attribute is removed from the data set, will it have a discrimination score of 0 which cannot happen due to the so-called red-lining effect. The red-lining effect is defined as other attributes in the model that are highly correlated with the sensitive attribute and thus result in the classifier correlating these new attributes to indirectly discriminate. In the example of risk score of prisoner defendants, removing ethnicity from the algorithm would not solve much if the algorithm has the address or postal code of the defendants

thus it could have biased if the defendant lives in a more black or more white concentrated area. The discrimination then would still be present just hidden better.

The first of approach Calders and Verwer study is where they modify the Naïve Bayes classifier. In the original Naïve Bayes model it uses conditional probability, which is defined as the probability of an event happening given that another event has already happened. In this algorithm, they would keep adding probability to the sensitive values $R_0$ with the positive class $Y_1$ and removing probability from sensitive values $R_1$ given the positive class. This will, however, have unwanted effects of always increasing or decreasing the number of positive labels assigned by the classifier which is dependent on whether favored sensitive values are more or less frequent in the data-set. This then caused them to change the Naïve Bayes model by changing $P(R|Y)$ into $P(Y|R)$, thus the joint distribution with all the other $A_1 \dots A_n$ attributes will now be:

$$P(Y, R, A_1 \dots A_n) = P(R)P(Y|R)P(A_1|Y) \dots P(A_n|Y)$$

.

They modify $P(Y|R)$ until the number of assigned positive labels and the number of positive labels in the data-set do not deviate by much. The paper also shows graphical models of this modified Naïve Bayes approach, where given the sensitive attribute R, it is used to calculate the probability of $Y$. Then with the calculated probability of Y and additional attributes $A_1 \dots A_n$, each additional attribute probability is calculated separately given the probability of $Y$ which was calculated from $R$. This is seen in Figure 2 in the first graph.

**Algorithm 1** Modifying Naive Bayes [1]

**Require:** a probabilistic classifier $M$ that uses distribution $P(Y|R)$ and a data-set $D$
**Ensure:** $M$ is modified such that it is (almost) non-discriminating, and the number of positive labels assigned by $M$ to items from $D$ is (almost) equal to the number of positive items in $D$.
Calculate the discrimination *disc* in the labels assigned by $M$ to $D$
**while** *disc* $> 0.0$ **do**
  *numpos* is the number of positive labels assigned by $M$ to $D$
  **if** *numpos* $<$ the number of positive labels in $D$ **then**
    $N(Y_1, R_0) = N(Y_1, R_0) + 0.01 * N(Y_0, R_1)$
    $N(Y_0, R_0) = N(Y_1, R_0) - 0.01 * N(Y_0, R_1)$
  **else**
    $N(Y_0, R_1) = N(Y_0, R_1) + 0.01 * N(Y_1, R_0)$
    $N(Y_1, R_1) = N(Y_0, R_1) - 0.01 * N(Y_1, R_0)$
  Update $M$ using the modified occurrence counts $N$ for $Y$ and $R$
  Calculate *disc*

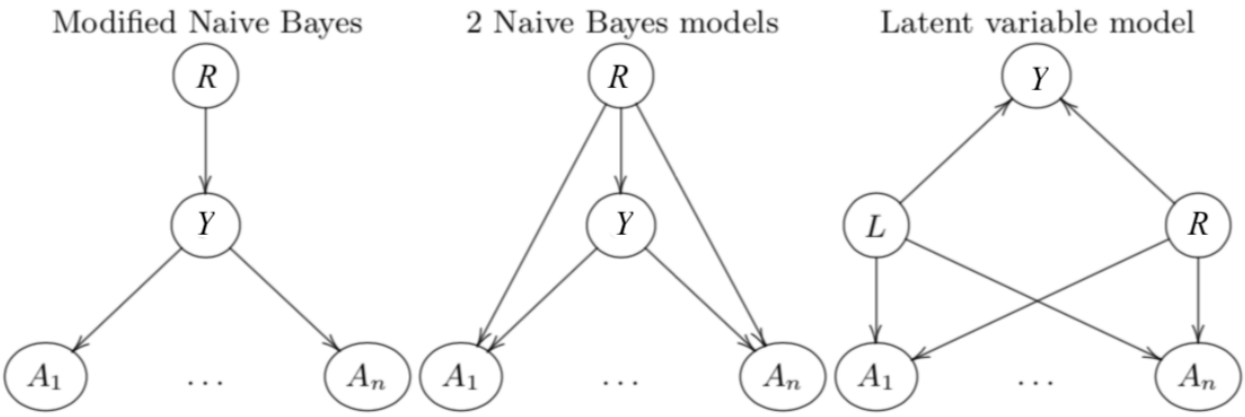



FIGURE 2. This figure shows the graphical models of the three Naive Bayes approaches for discrimination-free classification from Calders and Verwer [1]. The first model is the regular modified Naive Bayes where the additional attributes are used to decide R. The second model is the modified Two Naive Bayes algorithm where two Naive Bayes algorithms are done two separate sets: $R_-$ and $R_+$ which then removes the fact that the additional attributes are used to decide the sensitive attribute R. The last model is the Latent model using the latent variable as seen latent variable L is independent of sensitive attribute R and Y is determined by using L and R.

Their algorithm, shown in Algorithm 1, strives to take out discrimination from the probabilistic classifier M and the number of positive labels assigned by M to items from data-set D is equal to the number of positive items in D. It does this by using the number of positive labels assigned by M to D and updating the occurrence counts N of $Y_1$, $Y_0$ and $R_0$ or $Y_1$, $Y_0$ and $R_1$ depending on if the number is less than or greater than the number of positive labels in D. The classifier M is then updated and the discrimination is again calculated until we hopefully get to 0.0 discrimination. Even though this algorithm takes out discrimination from the classifier as said it does not actively avoid the red-lining effect and thus the decision is not necessarily independent from the correlated additional attributes.

The second approach they use is the two Naïve Bayes models which avoids the dependence on the additional attributes by removing the correlation between $R$ and $A$ from the data-set which was done in the first approach. This can be done by removing the additional attributes from the data-set, however, the consequences would be a big loss in accuracy due to the reduction in the number of attributes. The approach the paper takes is not removing the additional attributes, but removing the fact that they can be used to decide

$R$. An easy way to achieve this would be to split the data-set into two sets: one for $R = 1$ and one for $R = 0$. Thus, the model $M_1$ will be learned from the $R_1$ set and the model $M_0$ will be learned from the $R_0$ set and the overall classifier chooses between $M_1$ or $M_0$. For example, in our research, we would learn a model for white defendants and a model for black defendants. In this approach, it would remove discrimination by modifying the probability $P(Y|R)$ using the same algorithm stated above. In Figure 2 of the two Naïve Bayes shows that the sensitive attribute R is split into two models and directly used with the additional attributes.

The last approach they use is the latent variable model which uses a latent variable $L$ to model the class labels and to discover the actual class labels that the data-set should have contained if it would be discrimination-free. For this model they assume the following:

1. $L$ is independent from $R$, i.e., the actual labels are discrimination-free. This assumption allows the model to focus on the overall discrimination and any other form of discrimination such as from the other attributes A.
2. $Y$ is determined by trying to take out discrimination from the $L$ label sets, $L_0$ and $L_1$, using $R$ uniformly at random. Every tuple has an equal chance of being discriminated, again independent of attributes $A_1..A_n$, and thus also independent of the probability of being assigned a positive label $P(L_1|A_1...A_n)$.

In this model, as with the two Naive Bayes model, removing that an attribute A can be used to decide R by splitting $P(A|Y)$ into $P(A|Y, R = 1)$ and $P(A|Y, R = 0)$ during classification.

Estimating the third model from a data-set is more difficult. In order to do so, find two groups (or clusters) of tuples: the ones that should have gotten a positive label $L_1$ and those that should have gotten a negative label $L_0$. An approach to find these clusters is to use the expectation maximization (EM) algorithm. Given a model $M$ with a latent attribute $L$, the goal of this algorithm is to set the parameters of M such that they maximize the likelihood of the data-set D. EM iteratively optimizes these settings given D (the M-step), then calculates the expected values of the L attribute given those settings (the E-step), and incorporates these into D. This is a greedy procedure that converges to a local optimum of the likelihood function.

However, there is a better manner than simply running EM in order to find the discrimination-free class labels $L_1$ and $L_0$ and hoping that that solution found corresponds to discrimination-free labels. Modifying the labels of tuples with favored sensitive values $R = 1$ and negative class labels $Y = 0$ is not optimal. The same holds for tuples with discriminated sensitive values $R = 0$ and positive class labels $Y = 1$. Modifying these only results in more discrimination, so the latent values of these tuples are fixed to be identical to the class labels in the data-set. These values are excluded from the E-step of the EM algorithm.

Another improvement over blindly applying EM is to incorporate prior knowledge of the distribution of $Y$ given $L$ and $R$ i.e. $P(Y|L,R)$. We can pre-compute this entire distribution using an example since the goal is to achieve discrimination.

Suppose we have a data-set consisting of 100 tuples, distributed according to the following frequency counts:

| | $R_1$ | $R_0$ |
|---|---|---|
| $Y_1$ | 40 | 20 |
| $Y_0$ | 10 | 30 |

There's discrimination because the ratio of tuples with $R_1$ that have a positive class label $Y_1(4/5)$ is larger than the ratio of tuples with $R_0$ that have the positive class(2/5). The ratio of tuples with $R_0$ that have a negative class label $Y_0$ (3/5) is larger than the ratio of tuples with $R_1$ with a negative class label $Y_0$ (1/5) which also shows discrimination. Initially, the distribution was set over $L$ to be equivalent to the distribution over $Y$, keeping the discrimination intact:

| | $R_1$ | | | $R_0$ | |
|---|---|---|---|---|---|
| | $L_1$ | $L_0$ | | $L_1$ | $L_0$ |
| $Y_1$ | 40 | 0 | $Y_1$ | 20 | 0 |
| $Y_0$ | 0 | 10 | $Y_0$ | 0 | 30 |

The goal here to have the same total number of positive L values, $L_1$, in both $R_1$ and $R_0$ while also having the same total number of negative L values, $L_0$, in $R_1$ and $R_0$. There is now an uneven distribution, thus in order to improve this situation we find a number $n$ that will hopefully balance the distribution. Taking this $n$ we subtract it from $L_1$ and $Y_1$ in $R_1$ and adding $n$ to $L_0$ and $Y_1$ in $R_1$. We then also subtract the same number $n$ from $L_0$ and $Y_0$ in $R_0$ and add $n$ to $L_1$ and $Y_0$ in $R_0$. In this case the $n$ that would give this even distribution, leading to zero discrimination, would equal to 10, resulting in the following distribution:

| | $R_1$ | | | $R_0$ | |
|---|---|---|---|---|---|
| | $L_1$ | $L_0$ | | $L_1$ | $L_0$ |
| $Y_1$ | 30 | 10 | $Y_1$ | 20 | 0 |
| $Y_0$ | 0 | 10 | $Y_0$ | 10 | 20 |

As the table above shows, there is now an even distribution of tuples that are $L_1$ in both $R_1$ and $R_0$ with the number of tuples being 30 and there is an even distribution of tuples that are $L_0$ in both $R_1$ and $R_0$ with the number of tuples being 20. These counts are then used to pre-compute the probability table $P(Y|L,R)$ in the latent variable model.

To test these models, they take artificial data to generate the class labels that should be assigned to the tuples when there is no discrimination, and real-world data, however with the real-world data there is no discrimination-free test set. In their result of using artificial data, they based their results in terms of discrimination scores and accuracy scores. In order to test on artificial data, they generate data and initialize the parameters of latent variable model $M$. One thing they put into consideration is that it is unlikely that the joint distribution of an attribute $A$ and the latent class $L$ is completely different for tuples with $R = 1$ than for those with $R = 0$. In terms of discrimination, the modified Naïve Bayes

| | $R$ included | | marginalizing over $R$ | |
|---|---|---|---|---|
| | discrimination | accuracy | discrimination | accuracy |
| NB | -0.003 | 0.813 | 0.286 | 0.818 |
| 2 models | -0.003 | 0.812 | 0.047 | 0.807 |
| EM | 0.000 | 0.773 | 0.081 | 0.739 |
| EM prior | 0.013 | 0.790 | 0.077 | 0.765 |
| EM stopped | -0.006 | 0.797 | 0.061 | 0.792 |
| EM prior stopped | -0.001 | 0.801 | 0.063 | 0.793 |

TABLE 2. Discrimination and accuracy values resulting from 10-fold cross-validation of all methods with and without marginalizing over $R$ on census income from the results of Calders and Verwer [1].

and the 2 NB model approaches did the best rather than the latent model. In terms of accuracy, the modified Naïve Bayes model scores less than the two Naïve Bayes model. They found out that the drop in accuracy is smaller than the drop in discrimination. They expected that if they model the latent variable $L$ where there is no dependence between $L$ and $R$, the maximum likelihood assignment of $L$ would also be without dependence, however, this turned out to be false.

They then test on a real world example which was on census income: a data-set containing numerical and categorical attributes that can be used to decide whether a new individual should be classified as having a high or a low income. This will hopefully be done with zero discrimination with respect to the gender attribute.

Their results for the real world case shows that both the modified Naive Bayes and 2 Naive Bayes models drop in accuracy is smaller than the drop in discrimination. However, this does not hold for the expectation maximization methods as they converge to a point that is worse in terms of accuracy and discrimination than the first time they reached zero discrimination. Another interesting observation is that, although the EM methods do not perform well in the end, they do start out good. It starts by fixing the latent values for females with a positive class label and males with a negative class label, randomizing the other latent values and using this set to estimate the latent variable model. Thus the latent variable model seems to perform well with respect to the other two approaches, only EM does not converge to the expected point. The results are summarized in Table 2 and show that both Naive Bayes methods perform very well when $R$ is included and using prior information in the EM method improves its performance. They then marginalize over $R$ which means that $R$ is unknown during the testing phase and shows the dependence of the classifiers on the $R$ attribute. The first interesting observation is that the modified Naive Bayes method obtains very high discrimination which is not as surprising as this model is almost identical to the to the standard Naive Bayes model when $R$ is removed from the training data-set. The second observation is that in the non-stopped EM methods, the accuracy drops, thus there is a high dependence on the $R$ attribute. The 2 Naive Bayes models method has the lowest dependence on $R$, resulting in only about 5% discrimination

if $R$ is removed which is surprising since it uses $R$ to split the data and then learn two separate models. From these observations the conclusion, of the experiments is that the 2 Naive Bayes models method performs best as it achieves high accuracy scores with zero discrimination, and has the smallest dependency on $R$.

3.4. **Zafar.** Instead of focusing on parity and outcome like the Calders and Verwer study, Zafar et al.[9] decide to focus on parity and error in their paper. They state that there are three different types of un-fairness in discrimination-aware tools and classifiers: disparate treatment, disparate impact, and disparate mistreatment. A decision-making process suffers from disparate mistreatment with respect to a sensitive attribute if the misclassification rates differ for the groups of people that have different values of that sensitive attribute. If in the training data there are items with positive and negative class labels that are not linearly separable, the classifier or algorithm will misclassify people. In their context, misclassification rates may differ for groups of people who have different values for the sensitive attributes and thus disparate misclassification may happen. Disparate treatment may arise when a decision-making system has different outputs for groups of people with the similar values of non-sensitive attributes but different values of sensitive attributes. Lastly, disparate impact may arise when a decision-making system has outputs that benefit or hurt a group of people that share a value of sensitive attribute more frequently than other groups of people.

Table 3 shows an example of three classifiers to see if the ground truth is true or false using the user attributes. It shows that classifier 1 suffers from disparate impact as the fraction of blacks and whites that were classified to recidivate are different (1.0 and .66 respectively). Classifier 2 and Classifier 3 suffer from disparate treatment as their decisions for White 3 and Black 1 to recidivate is different even though they have the same values for non-sensitive attributes. It is seen that Classifier 1 and Classifier 2 suffer from disparate mistreatment as Classifier 1 has different false negative rates for black and white defendants and Classifier 2 has both different false negative and false positive rates for black and white defendants.

In formalizing notions of fairness, Zafar et al. looked at creating a classifier that avoids disparate treatment, disparate impact and disparate mistreatment and approaches to achieve this. A binary classifier does not suffer from disparate treatment if the probability that the classifier outputs a value in the class labels in $\{0, 1\}$, positive or negative classes, given a feature vector does not change after observing a sensitive feature i.e.

$$P(S|x, R) = P(S|x)$$

A binary classifier does not suffer from disparate impact if the probability that the classifier assigns a group to the positive class is the same for both values of the sensitive feature which is defined as being in $\{0, 1\}$ i.e.

$$P(S = 1|R = 0) = P(S = 1|R = 1)$$

| User Attributes | | |
|---|---|---|
| Sensitive | Non-Sensitive | |
| Race | Previous non-violent crime | Previous violent crime |
| White 1 | 1 | 1 |
| White 2 | 0 | 1 |
| White 3 | 0 | 0 |
| Black 1 | 0 | 0 |
| Black 2 | 1 | 1 |
| Black 3 | 1 | 0 |

| Ground Truth (recidivated) |
|---|
| Yes |
| No |
| No |
| Yes |
| Yes |
| No |

| Classifier's Decision for low or high risk | | |
|---|---|---|
| $C_1$ | $C_2$ | $C_3$ |
| 1 | 1 | 1 |
| 1 | 1 | 0 |
| 0 | 0 | 0 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 0 |

TABLE 3. Decisions of three fictitious classifiers ($C_1$, $C_2$ and $C_3$) whether the six defendants are high (1) or low (0) risk to recidivate after they are released. Race is the sensitive attribute, whereas the other two attributes (previous non-violent crime and previous violent crime) are non-sensitive. Ground truth of whether the defendant actually recidivated is also shown.

As seen disparate impact is much like statistical parity talked about in Chouldechova, however, disparate impact does not depend on a threshold score only positive, 1, or negative, 0, class labels. A binary classifier does not suffer from disparate mistreatment if the misclassification rates for different groups of people having different values of the sensitive feature R are the same. Misclassification rates are equal to the fraction of false positive and false positive rates i.e. as fractions over the class distribution in the ground truth labels i.e. as false omission and false discovery rates. Disparate misclassification can be defined with these definitions: *overall misclassification rate (OMR):*

$$P(S \neq Y|R = 0) = P(S \neq Y|R = 1)$$

*false positive rate (FPR):*

$$P(S \neq Y|R = 0, Y = 0) = P(S \neq Y|R = 1, Y = 0)$$

*false negative rate (FNR):*

$$P(S \neq Y|R = 0, Y = 1) = P(S \neq Y|R = 1, Y = 1)$$

*false omission rate (FOR):*

$$P(S \neq Y|R = 0, S = -1) = P(S \neq Y|R = 1, S = -1)$$

*false discovery rates (FDR):*

$$P(S \neq Y|R = 0, S = 1) = P(S \neq Y|R = 1, S = 1)$$

Chouldechova and Kleinberg et al. show that when the fraction of users with positive class labels differ between members of different sensitive attribute value groups, it is impossible to construct classifiers that are equally well-calibrated and also satisfy the equal false positive and false negative rates. These results suggest that satisfying all five criterion of disparate mistreatment simultaneously is impossible when the underlying distribution of data is different for different groups.

Zafar et al. then try to train a classifier that does not suffer from disparate mistreatment. These classifiers generally learn the optimal decision boundary by minimizing a convex loss $L(\theta)$. The convexity of $L(\theta)$ ensures that a global optimum can be found efficiently. They were able to find a linear program that minimizes the loss function:

$$\begin{aligned} \text{minimize} \quad & -\sum_{(x,Y)\in D} log\ p(Y_i|x_i, \theta) \\ \text{subject to} \quad & -\frac{N_1}{N}\sum_{(x,Y)\in D_0} g_\theta(Y, x) + \frac{N_0}{N}\sum_{(x,Y)\in D_1} g_\theta(Y, x) \leq c \\ & -\frac{N_1}{N}\sum_{(x,Y)\in D_0} g_\theta(Y, x) + \frac{N_0}{N}\sum_{(x,Y)\in D_1} g_\theta(Y, x) \geq -c \end{aligned}$$

This linear program shows that in order to minimize the convex loss $L(\theta)$ which is equal to $-\sum_{(x,Y)\in D} log\ p(Y_i|x_i, \theta)$ and without disparate mistreatment, the two conditions need to be met. The linear program is a Disciplined Convex-Concave Program (DCCP) for any convex loss $L(\theta)$, and can be efficiently solved using well-known heuristics. In order to understand these conditions, $x$ is a user vector $\in \mathbb{R}^d$ and class labels $Y \in \{0, 1\}$. $D$ is training dataset and $D_0$ and $D_1$ are subsets of $D$. The constant number $\frac{1}{N}$ can be dropped while $N_0 = |D_0|$ and $N_1 = |D_1|$. The function $g_\theta(Y, x)$ is defined as:

$$g_\theta(Y, x) = min(0, Y d_\theta(x)),$$

$$g_\theta(Y, x) = min(0, \frac{1 - Y}{2} Y d_\theta(x)), or$$

$$g_\theta(Y, x) = min(0, \frac{1 + Y}{2} Y d_\theta(x))$$

In linear models for classification the decision boundary is simply the hyperplane defined by $\theta^T x = 0$, therefore, $d_\theta(x) = \theta^T x$. The covariance threshold $c \in \mathbb{R}^+$ controls how adherent to disparate mistreatment the boundary should be. The above linear program ensures that the classifier chooses the optimal decision boundary within the space of fair boundaries specified by the constraints. It provides the flexibility to remove disparate treatment as it removes disparate mistreatment. Since their formulation does not require the sensitive attribute information at decision time, by keeping the features $x$ disjoint from sensitive attribute $R$, one can remove disparate mistreatment and disparate treatment simultaneously.

In order to evaluate the effectiveness of their classifiers in controlling disparate mistreatment, they conducted experiments on both synthetic and real world datasets. They first define how they quantify the disparate mistreatment incurred by as a classifier:

$$D_{FPR} = P(S \neq y | R = 0, Y = 0) - P(S \neq y | R = 1, Y = 0),$$
$$D_{FNR} = P(S \neq y | R = 0, Y = 1) - P(S \neq y | R = 1, Y = 1)$$

where the closer the values of $D_{FPR}$ and $D_{FNR}$ are to 0, the lower the degree of disparate mistreatment.

When testing on synthetic data, the first scenario tested is with the classifier being unfair in terms of only false positive rate or false negative rate. In their results they analyze (a) the relation between decision-boundary covariance and the false positive rates for both sensitive attribute values; (b) the trade-off between accuracy and fairness; and (c) the decision boundaries for both the unconstrained classifier and the fair constrained classifier. The found that 1) as the fairness constraint value $c = mc^*$ goes to zero, the false positive rates for both groups ($R = 0$ and $R = 1$) converge, and hence, the outcomes of the classifier become fairer i.e. $D_{FPR} \to 0$, while $D_{FNR}$ remains close to zero and 2) ensuring lower values of disparate mistreatment leads to a larger drop in accuracy. Another algorithm they used to analyze performance was called *Baseline*. This approach tries to remove disparate mistreatment by having different penalties for misclassified data points with different sensitive attrbute values during training phase. It achieves this in two steps. First, it trains an un-fair classifier minimizing a loss function over the training data. Next, it takes the set of misclassified data points from the sensitive attribute valye group that presents the higher error rate. Take for example if one wants to remove disparate mistreatment with respect to false positive rate and $D_{FPR} > 0$, this approach will select the set of misclassified data points in the training set having $R = 0$ and $Y = 0$. Next, it iteratively re-trains the classifier with increasingly higher penalties on this set of data points until a certain fairness decision boundary is achieved in the training set. This approach does not use sensitive attribute information while making decisions thus will not suffer from disparate treatment like Zafar et al.'s original method.

The next step is to test for when the classifier is unfair in terms of both false positive rate and false negative rate. They first observed the case when $D_{FPR}$ and $D_{FNR}$ have opposite signs, i.e. false positive rate for one group is higher than the other, while the

false negative rate for the same group is lower. In their results they found that first removing disparate mistreatment on only false positive rate causes a rotation in the decision boundary to move previously misclassified examples with $R = 1$ into the negative class, decreasing their false positive rate. However, in the process it also moves previously well-classified examples with $R = 1$ into the negative class, increasing their false negative rate. As a consequence, controlling disparate mistreatment on false positive rate also removes disparate mistreatment on false negative rate. The second case they observed is when $D_{FPR}$ and $D_{FNR}$ have the same signs i.e. both false positive as well as false negative rate are higher for a certain sensitive attribute value group. The results show that first controlling disparate mistreatment for only false positive rate, leads to a minor drop in accuracy, but can worsen the disparate mistreatment on false negative rate. This similarly happens when controlling disparate mistreatment with respect to only false positive rate. As a consequence, controlling for both types of disparate mistreatment simultaneously brings $D_{FPR}$ and $D_{FNR}$ close to zero, but causes a large drop in accuracy.

They then use their classifier in order to see if disparate mistreatment reduces accuracy. Their results show that similar to their results from their synthetic results, for all three methods, FPR, FNR and both constraints, controlling for disparate mistreatment on false positive rate (false negative rate) also helps decrease disparate mistreatment on false negative rate (false positive rate). Zafar et el.'s and Hardt et al.'s algorithm are also able to achieve similar accuracy for a given level of fairness. Their results showed that Baseline does the worst in accuracy, about 20% less than Zafar et al.'s or Hardt et al.'s methods and does not remove disparate mistreatment completely, i.e. it does not achieve zero $D_{FPR}$ or/and $D_{FNR}$ in any of the cases. They also discovered that their method does not completely remove disparate mistreatment. This shows that their method can suffer from reduced performance on small datasets. They acknowledge that with large training datasets, they expect more reliable estimates of covariance and a better performance from their method.

In conclusion, Zafar et al's method to avoid disparate mistreatment and disparate treatment in recidivism assessment tools is successful as they were able to do it simultaneously. Disparate mistreatment is avoided by using fairness constraints. Disparate treatment is avoided by ensuring that sensitive attribute information is not used while making decisions i.e. by keeping user feature vectors ($x$) and the sensitive features ($R$) disjoint. This feature might be useful in scenarios when the sensitive attribute information is not available such as in our dataset. In their future work, they acknowledge that they would include other measures of disparate mistreatment into their fair classifier formulation such as false discovery and false omission rates. They did not include these factors in their current method as including false discovery and false omission rates are a non-trivial task due to computational complexities involved.

3.5. **Hardt et al.** Zafar et al. analyze their classifier's performance in comparison to Hardt et al. [5] and show that Hardt et al.'s classifier has a better performance in accuracy, however it suffers from disparate treatment. Thus, it makes sense to take some time to talk about Hardt et al. and their classifier for a discrimination-free method. Hardt et al. show in their paper the errors in discrimination-aware classifiers against protected attributes, much like Zafar et al. They first talk about common conceptions of nondiscrimination such as redundant encodings, which are ways of predicting protected attributes from other features much like red-lining effect as talked about in Calders and Verwer's paper, and demographic parity. Demographic parity requires that a decision be independent of the protected attribute. They argue that demographic parity is flawed in that it does not ensure fairness and it cripples the utility that is hoped to be achieved. They consider non-discrimination from the perspective of supervised learning, where the goal is to predict a true outcome $Y$ from features $X$ based on labeled training data while ensuring the prediction is not discriminatory with respect to a protected attribute $R$. They then further define different criteria for a classifier to not be discriminatory. First is that a predictor $\widehat{Y}$ satisfies equalized odds with respect to protected attribute $R$ and outcome $Y$, if $\widehat{Y}$ and $R$ are independent conditional on $Y$ which is equivalent to:

$$P(\widehat{Y} = 1|R = 0, Y = y) = P(\widehat{Y} = 1|R = 1, Y = y), y \in \{0, 1\}$$

For the outcome $y = 1$, the constraint requires that $\widehat{Y}$ has equal true positive rates across the two sensitive attributes $R = 0$ and $R = 1$. For $y = 0$, the constraint equalizes false positive rates. Equalized odds enforces both equal bias and equal accuracy in all demographics, punishing models that perform well only on the majority. Unlike demographic parity, equalized odds allows the predictor to depend on the protected attribute but only through the target variable $Y$.

Next is the notion of equal opportunity which a binary predictor $\widehat{Y}$ satisfies equal opportunity with respect to $R$ and $Y$ if

$$P(\widehat{Y} = 1|R = 0, Y = 1) = P(\widehat{Y} = 1|R = 1, Y = 1)$$

As seen this definition of equal opportunity is similar to Zafar et al's notion of disparate treatment as they both show that a classifier should be assigned to the positive class for both values of the sensitive attributes.

The third definition is the notion of oblivious: predictor $\widehat{Y}$ or score $S$ is said to be oblivious if it only depends on the joint distribution of $(Y, R, \widehat{Y})$ or $(Y, R, S)$, respectively. A real-valued predictive score $S = f(X, R)$ where higher values of $S$ correspond to greater likelihood of $Y = 1$ and thus a bias toward $\widehat{Y} = 1$. A binary classifier $\widehat{Y}$ can be obtained by thresholding the score, i.e. setting $\widehat{Y} = \mathbb{I}\{S > t\}$ for some threshold t. This is also seen in Chouldechova's definition of high-risk threshold where the predictor will be classified as high-risk above this threshold and low-risk below this threshold. A score $S$ satisfies equalized odds if $S$ is independent of $R$ given $Y$. If a score obeys equalized odds, then any thresholding of it also obeys equalized odds. As a consequence of being oblivious, all the

information needed to verify the definitions stated is contained in the joint distribution of the predictor, protected group and outcomes, $(\widehat{Y}, R, Y)$. Thus it will be assumed that the joint distribution of $(\widehat{Y}, R, Y)$ is known.

They then explain how to obtain an equalized odds or equal opportunity predictor $\widehat{Y}$ from a learned binary predictor $\widehat{Y}$ or score $S$. They do not require changing the training process, as this might introduce additional complexity, but rather only a post-learning step. Instead, a non-discriminating predictor which is derived from $\widehat{Y}$ or $S$ will be constructed. A predictor $\widehat{Y}$is derived from a random variable $S$ and the protected attribute $R$ if it a possibly randomized function of the random variables $(S, R)$ alone. In particular, $\widehat{Y}$ is independent of $X$ conditional on $(S, R)$. The definition asks that the value of a derived predictor $\widehat{Y}$ should only depend on $S$ and the protected attribute, though it may introduce additional randomness.

In order to obtain a good predictor satisfying equalized odds, a loss function $l : \{0, 1\}^2 \to \mathbb{R}$ is done. The loss function takes a pair of labels and returns a real number $l(\widehat{y}, y) \in \mathbb{R}$ which indicates the loss of predicting $\widehat{y}$ when the correct label is y. The goal is then to design derived predictors $\widehat{Y}$ that minimizes the expected loss $\mathbb{E}l(\widehat{Y}, Y)$ subject to one of their definitions.

3.5.1. *Deriving from a binary predictor.* In designing a derived predictor from binary $\widehat{Y}$ and $R$, four parameters are set: the conditional probabilities $p_{ya} = P(\tilde{Y} = 1|\widehat{Y} = a, R = a)$. The four parameters, $p = (p_{00}, p_{01}, p_{10}, p_{11})$ specify the derived predictor $\tilde{Y}_p$. To check whether $\tilde{Y}_p$ satisfies equalized odds we need to verify the two equalities specified in the definition of equalized odds, for both values of $y$:

$$\gamma_a(\tilde{Y}) = (P(\tilde{Y} = 1|R = a, Y = 0), P(\tilde{Y} = 1|R = a, Y = 1))$$

The components of $\gamma_a(\tilde{Y})$ are the false positive rate and the true positive rate within the demographic R = a. $\tilde{Y}$ satisfies equalized odds iff $\gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y})$.

3.5.2. *Deriving from a score function.* Another way of deriving a binary predictor from a score $S$ that does not use the protected attribute would be to threshold it, i.e. using $\widehat{Y} = \mathbb{I}\{S > t\}$. If $S$ satisfied equalized odds, then so will such a predictor, and the optimal threshold should be chosen to balance false and true positive rates so as to minimize the expected loss. When $S$ does not already satisfy equalized odds, we might need to use different thresholds for different values of $R$ (different protected groups), i.e. $\tilde{Y} = \mathbb{I}\{S > t_R\}$, however, randomness may be introduced as this might not be sufficient.

Hardt et al. wanted to accomplish two important criteria in their fairness measure: first to remedy the main conceptual shortcomings of demographic parity and second to align with the central goal of supervised machine learning, that is, to build higher accuracy classifiers. Their notion requires access to observed outcomes such as if a defendant committed the crime they were tried for in our setting. This is the same requirement that supervised learning generally has, however, having access to reliable "labeled data" is not

always possible. Moreover, the measurement of the target variable might in itself be unreliable or biased. Domain-specific analysis is required in defining and collecting a reliable target variable. Requiring equalized odds creates an incentive structure for building the predictor that aligns well with achieving fairness. Achieving better prediction with equalized odds requires collecting features that more directly capture the target, unrelated to its correlation with the protected attribute. An important feature of our notion is that it can be achieved via a simple and efficient post-processing step.

3.6. **Summary of Literature Review.** Starting off this discussion were Kleinberg et al and Chouldechova in order to understand what it means to be a risk assessment instrument. Kleinberg et al give an introduction on the variables needed to for a risk classifier that hopefully does not have any discrimination such as calibration within groups, balance for the negative class and balance for the positive class as well as discussion of different fields that use risk assessment instruments. Chouldechova also talks about the requirements to be a risk assessment instrument, however focussing on COMPAS which was the algorithm ProPublica discussed in their paper. Using Chouldechova, we can understand what is wrong with COMPAS and how to hopefully give less discrimination in the score. Chouldechova agreed with Kleinberg et al. that in order for the algorithm to not have discrimination is calibration and error rate balance. She proposes that the algorithm also needs to satisfy predictive parity and statistical parity, however she mostly focuses on the former. She is able to show how disparate impact can result from the use of a recidivism instrument that is known to satisfy predictive parity. In our analysis to see if rearrests based on convictions gives less discrimination than recidivism data used in ProPublica, we will have to see if false positive rates and false negative rates are different and using them to measure disparate parity. We will also have to see if in our data the minimum and maximum time that the defendant is expected to have for the crime in order to accurately measure disparate impact. We also have to recognize that we have some control over the PPV and error rates when using her classifier and thus seeing if we can change these factors to satisfy predictive parity and thus disparate impact.

Calders and Verwer's was the first introduction into the discrimination-free classifiers studies in 2010 in comparison to Zafar et al and Hardt et al which were released in 2016. Understanding their three classifiers gives us an understanding of the earlier algorithms and how these classifiers have evolved using Zafar et al and Hardt et al. as examples of this evolution. While Kleinberg and Chouldechova discuss how to remove machine bias in their classifiers, they do not use discrimination scores as their measure of success. In Calders and Verwer's classifiers they use the dataset, a binary class attribute and binary sensitive attribute in order to define discrimination and do not take account a label if action done was done or not, while Zafar does take this into account. Another difference between Calders and Verwer's and Zafar et al.'s classifier(s) is that the former assumes the false positive and false negative rates are equal while the latter does not and uses these rates to calculate disparate mistreatment. In Zafar et al's results of COMPAS with their classifier, due to the small dataset, there was more discrimination found and less accurate

in comparison to their synthetic data analysis. They used the ground truth of whether or not the defendants studied actually recidivated within two years after the screening while the ground truth we will be using if the defendant was actually found guilty for the crime they were arrested for. Even though the ground truth is more accurate, our dataset may be smaller than ProPublica's study and thus not enough for Zafar et al's classifier to be trained fairly which could lead to more discrimination and less accuracy. Calders and Verwer did not use this ground truth label in their classifiers. In their analysis of their three classifiers, they determined the discrimination score and accuracy when the classifier is dependent on the sensitive attribute or not. They found that their 2 Naive Bayes models performed the best. When looking at the results of both Zafar et al. and Calders and Verwer, it seems that Calders' classifiers perform better as there is about the same discrimination found the accuracy of the 2 Naive Bayes classifiers is higher.

Zafar et al compare their classifier with Hardt et al's classifier. They also have another method that uses sensitive attribute to avoid disparate mistreatment only. The claim that when the sensitive attribute information is used for decision making, resulting in disparate treatment. Hardt et al's classifier operates by post-processing the outcomes of an unfair classifier and using different decision thresholds for different sensitive attribute value groups to achieve fairness. It needs the sensitive attribute information while making decisions and hence cannot avoid disparate treatment. For synthetic data, both Hardt et al's and Zafar et al's methods that use sensitive feature information present the best performance in terms of accuracy, however they both suffer from disparate treatment. Zafar et al's method without the use of sensitive information to simultaneously remove disparate mistreatment and disparate treatment does so with further accuracy drop of only 5% with respect to the above two methods that cause disparate treatment. Thus, Zafar et al's method without sensitive attributes and Hardt et al's method achieve the same performance when making use of the same information in the data, however Zafar et al's method simultaneously remove both disparate mistreatment and disparate treatment at a small additional cost in terms of accuracy. Thus comparing two methods that are the most current thus far Zafar et al's method results in less discrimination for the cost of accuracy even though the difference in accuracy is not substantial. In comparison between Hardt et al and Zafar et al's methods for COMPAS, Hardt et al. is able to achieve both zero $D_{FPR}$ and $D_{FNR}$ while controlling for disparate mistreatment on both false positive and false negative rates even though there is a considerable drop in terms of accuracy. Since Hardt et al.'s method operates on a data of much smaller dimensionality, it is not expected to suffer as much from the small size of the dataset as compared to Zafar et al.'s method.

## 4. Data Analysis and Results

Data collection was done by another student the previous semester [10]. With this data, I will look into the trends and analyze if there are any interesting trends between re-arrest and convictions. I will be seeing the difference in False Positive and False Negative Rate using different measures of recidivism, similar to the ProPublica data analysis. The student

found out when recreating recidivism and violent recidivism analysis of ProPublica using convictions instead of re-arrest, there was not much a difference.

4.1. **Intersectionality.** I will first be looking at the intersectionality of the data such as if there are interesting trends between race and sex when analyzing re-arrest vs conviction. I will first be looking at race and sex such as African-American and female, Caucasian and female, African-American and male, and Caucasian and male. The data used is the cleaned up version of the ProPublica dataset that Jai N. [10] used to analyze re-arrest vs. conviction false positive and false negative rates.

Below is the data taken from Jai for re-arrest and conviction error rates for just Caucasian and African-Americans in order to compare error rates for different intersectional data observed later.

| | White | African-American |
|---|---|---|
| Overall | 2103 | 3175 |
| Re-arrest | 822 | 1661 |
| Re-arrest Rate | 39.08% | 52.31% |
| Convictions | 114 | 347 |
| Conviction Rate | 5.42% | 10.93% |

TABLE 4. Dataset Demographic

| | White | African American |
|---|---|---|
| Labeled Higher Risk and Re-offened | 20.45% | 39.31% |
| Labeled Lower Risk and Didn't Re-offend | 45.79% | 25.86% |
| Labeled Higher Risk, but Didn't Re-offend | 13.41% | 18.30% |
| Labeled Lower Risk, but Did Re-offend | 19.40% | 16.54% |

TABLE 5. Re-arrest Error Rate by Race

| | White | African American |
|---|---|---|
| Labeled Higher Risk and Re-offened | 3.90% | 9.32% |
| Labeled Lower Risk and Didn't Re-offend | 65.38% | 40.79% |
| Labeled Higher Risk, but Didn't Re-offend | 29.20% | 48.28% |
| Labeled Lower Risk, but Did Re-offend | 1.52% | 1.61% |

TABLE 6. Conviction Error Rate by Race

I first ran experiments for re-arrest and conviction rates using race and sex data. The number of people who were re-arrested and the number who were convicted of a crime within two years of their COMPAS screening are shown below, split by race and female.

| | White Female | African-American Female |
|---|---|---|
| Overall | 482 | 549 |
| Re-arrest | 170 | 203 |
| Re-arrest Rate | 35.27% | 36.98% |
| Convictions | 18 | 26 |
| Conviction Rate | 3.73% | 4.74% |

TABLE 7. Dataset Demographic Female and Race

Below is a comparison of error rates with the first one comparing re-arrest error rates and the table below comparing conviction error rates.

| | White Female | African American Female |
|---|---|---|
| Labeled Higher Risk and Re-offened | 19.71% | 27.14% |
| Labeled Lower Risk and Didn't Re-offend | 44.81% | 38.25% |
| Labeled Higher Risk, but Didn't Re-offend | 18.46% | 22.40% |
| Labeled Lower Risk, but Did Re-offend | 17.01% | 12.20% |

TABLE 8. Re-arrest Error Rate by Race and Female

Below is the data for Race and female conviction rates.

| | White Female | African American Female |
|---|---|---|
| Labeled Higher Risk and Re-offened | 2.90% | 3.83% |
| Labeled Lower Risk and Didn't Re-offend | 61.00% | 49.54% |
| Labeled Higher Risk, but Didn't Re-offend | 35.27% | 45.72% |
| Labeled Lower Risk, but Did Re-offend | 0.83% | 0.91% |

TABLE 9. Conviction Error Rate by Race and Female

As seen in tables 8 and 9, the False Negative Rates for convictions are about twice as high and the True Positive Rates for convictions are about 1.5 times the rates for re-arrest for both White and African-American females. The True Positive and False Negative Rates go down to the point that they are about equal. Running a logistic regression and filtering the sex to be only female, African-Americans were 1.25 as likely to get a high score as

White defendants for re-arrest while using relative rates for convictions African-American defendants were 1.21 as likely to get a high score as White defendants.

Next is the data for Race and male re-arrest rates.

| | White Male | African-American Male |
|---|---|---|
| Overall | 1621 | 2626 |
| Re-arrest | 652 | 1458 |
| Re-arrest Rate | 40.22% | 55.52% |
| Convictions | 96 | 321 |
| Conviction Rate | 5.92% | 12.22% |

TABLE 10. Dataset Demographic Male and Race

Females had similar re-arrest and conviction rates depending on race, while males had more of a discrepancy. African-American males conviction rates were about twice a much as White male's conviction rates; African-American male's re-arrest rates were about 15% more than White male's re-arrest rates.

| | White Male | African American Male |
|---|---|---|
| Labeled Higher Risk and Re-offened | 20.67% | 41.85% |
| Labeled Lower Risk and Didn't Re-offend | 46.08% | 23.27% |
| Labeled Higher Risk, but Didn't Re-offend | 10.92% | 17.44% |
| Labeled Lower Risk, but Did Re-offend | 22.33% | 17.44% |

TABLE 11. Re-arrest Error Rate by Race and Male

It is interesting that False Positive and False Negative Rates for African-American males are equal to each other while False Negative Rate for White male is more than twice as the False Positive Rate for White males. Next is the data for Race and male conviction rates.

| | White Male | African American Male |
|---|---|---|
| Labeled Higher Risk and Re-offened | 4.19% | 10.47% |
| Labeled Lower Risk and Didn't Re-offend | 66.69% | 38.96% |
| Labeled Higher Risk, but Didn't Re-offend | 27.39% | 48.82% |
| Labeled Lower Risk, but Did Re-offend | 1.73% | 1.75% |

TABLE 12. Conviction Error Rate by Race and Male

Males yield similar results in error rates in comparison between re-arrest and convictions as females. Both True Negative and False Positive Rates go up from re-arrest to conviction, where False Positive Rates for convictions are twice as much as the rates for re-arrest. Again, True Positive and False Negative rates go down from re-arrest to convictions with

False Negative Rates being almost the same between White and African-American males. Taking a logistic regression with sex being consistent with being only males, African-Americans were 1.80 times more likely to have a higher score than White defendants when accounting for relative rates of convictions, while African-Americans were 1.89 times more likely to have a higher score than White defendants when accounting for relative rates of re-arrest.

This data showed that error rates for intersectionality and re-arrest is similar to ProPublica's findings for re-arrest and Jai's findings for convictions in that African-American females and males are more likely to have a higher False Positive Rate than White females and males while White females and males are more likely to have a higher False Negative Rate than African-American females and males. Conviction data shows that there is an even more difference between African-American and White False Positive Rates as the differences are much higher than the differences in re-arrest for all the intersectionality data, while the False Negative Rates for the conviction data for Caucasian and African-Americans were almost similar to each other.

4.2. **Fairness Algorithm Comparisons.** The *fairness* package was created by Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth [11]. It is meant to facilitate the benchmarking of fairness aware machine learning. Figure 3 shows the stages in how each algorithm is run in their fairness-

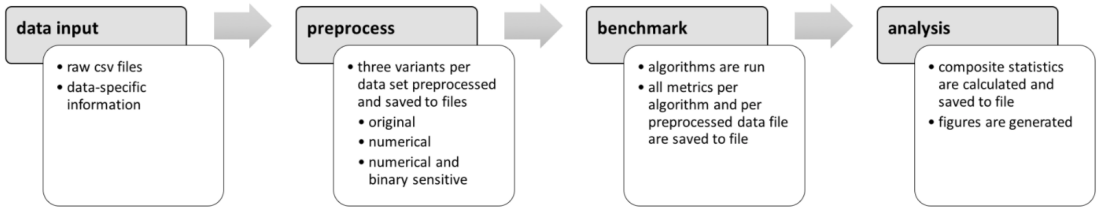



FIGURE 3. The stages of the fairness-aware benchmarking program: data input, preprocessing, benchmarking, and analysis. Intermediate files are saved at each stage of the pipeline to ensure reproducibility. [11]

aware benchmarking program. Once the package was installed, in order to compare the performance of Zafar et al.'s and Calders' algorithms between re-arrest and convictions, the already installed PropublicaRecidivism dataset was re-configured in order to have the *class_attr* variable in the dataset to be *two_year_conv* instead of *two_year_recid*. Once this was done, this new dataset was run with different algorithms, ZafarFairness, ZafarBaseline and Calders, for sex, race and both race and sex. Their first preprocessing step was to modify the input data such as removing features that should not be used for classification, removing or imputing any missing data and potentially removing items or adding derived features. In order to allow the analysis of fairness based on multiple sensitive attributes, such as ensuring fairness based on someone's race and sex, they also added a combined

sensitive attribute, which will be called sex-race in this analysis to both re-arrest and convictions dataset.

The algorithms that we are analyzing, Zafar et al. and Calders, require that the sensitive attributes be binary (such as "White" and "not White" instead of handling multiple racial categories) thus we will be using numerical+binary version of the data where the given privileged group is 1 and all other values to be 0.

Each algorithm was run over ten random 2/3 training set and 1/3 testing set splits and the result on each split is shown as a single point on the plot. With this data, I first produced a plot to show accuracy vs. (binarized) disparate impact for race, sex and both as seen below. This measurement is based on base rates and is defined as:

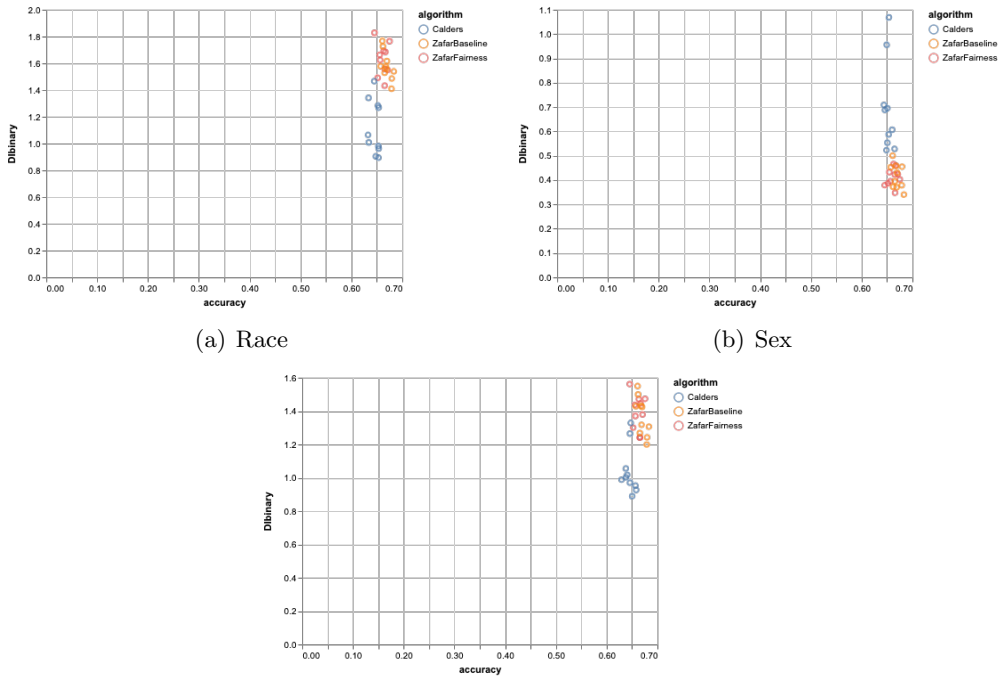
$$\frac{P[S = 1|R \neq 1]}{P[S = 1|R = 1]}$$

where $R$ is the sensitive attribute and $S$ is the predicted outcomes of the algorithm. Each classifier is considered fair if the fraction is between 0.8 and 1.2, if the fraction is higher than 1 then there is higher benefit for the unprivileged group and if less than 1 then there is higher benefit for of the privileged group. There are two variants of DI, binary and average, however I looked at binary, DIbinary, because all unprivileged classes are grouped together into a single value for non-White defendants that is compared as a group to the privileged class.

In order to see if Zafar et al.'s (ZafarFairness) and Calders' algorithms are fair relative to other algorithms for re-arrest and convictions, they were plotted against the ZafarBaseline algorithm. As seen above in the literature review, ZafarBaseline's algorithm tries to achieve fairness introducing different penalties for misclassified data points with different sensitive attribute values during training phase. [9]

Comparing the figures in Figure 4 and Figure 5 , the accuracy from using re-arrest data to conviction data jumped from an average of .65 to an average of .90. This shows that using conviction data creates more accuracy in fairness analysis. As seen in re-arrest data, disparate impact scale is similar for race, sex and sex-race, where the highest is near 2.0. For disparate impact, fairness is between 0.8 and 1.2 as it is the ratio rate for the unprivileged group to that of the privileged group. Thus, for re-arrest, Calders is fairer than Zafar et al.'s algorithms where Zafar et al.'s algorithms are more biased towards the Black defendants. Meanwhile, there is a higher benefit for Male defendants for both Calders and Zafar. Then for sex-race, the data is predictable to what is seen for race and sex where Calders is fairer for sex-race than Zafar, where Zafar has a higher benefit for Male and White defendants.

There is a big change in the scale when going from re-arrest to convictions for disparate impact where convictions is seen in Figure 5. The points are scattered all along from the scale 1 to 12 for race, 0 to 3 for sex and 1 to 9 for sex-race. Due to the fact that for each point is a random $\frac{2}{3} : \frac{1}{3}$, the 2/3 split used for the training may favor one group a lot more favorably than another since from the demographics shown in tables 4, 7 and 10, show that there were only about 5-10% of defendants that were convicted so each subpopulation consisted of a disproportionate amount of more non-convicted defendants than convicted




(a) Race

(b) Sex

(c) Sex-Race

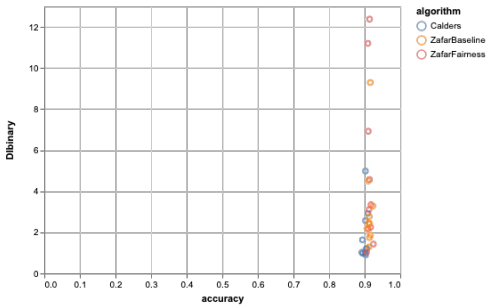
FIGURE 4. Rearrest DIbinary vs Accuracy.

defendants. Thus the ratios of people from race, sex and sex-race can be skewed to one type of the subpopulation. This can especially be seen when looking at race and sex-race where ZafarFairness and ZafarBaseline have the greatest range of disparate impact as there is a smaller ratio between the number of training examples and number of learnable features it hinders the estimate of misclassification covariance.

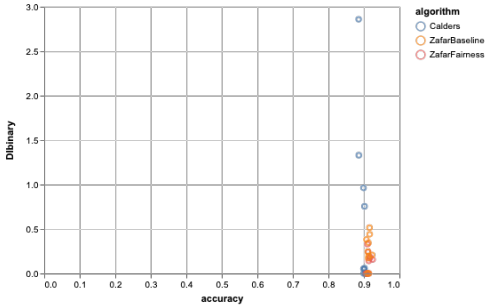
I then looked at Calders and Verwer's definition of non-discrimination of a dataset through their discrimination score. The CV fairness measure is based on Calders and Verwer's definition of fairness where the difference is taken between the privileged group and the unprivileged group. It is defined as:

$$1 - (P[S = 1|R = 1] - P[S = 1|R \neq 1])$$

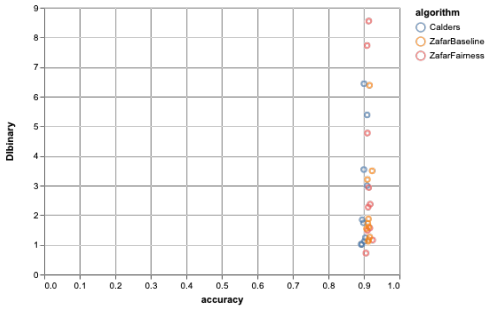
. This measure is the same as DI, however the difference is taken instead of the ratio. Fairness is between 0.8 and 1.2 again as this difference is subtracted from 1 thus if discrimination score is 0 then the total fairness is 0, thus CV = 1 is the most ideal.




(a) Race

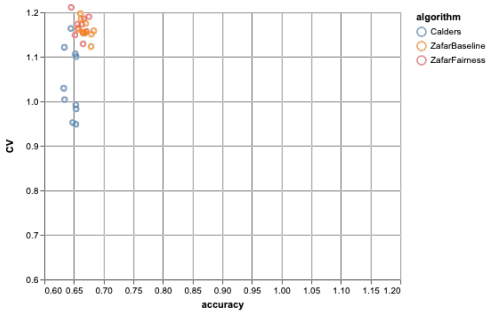



(b) Sex

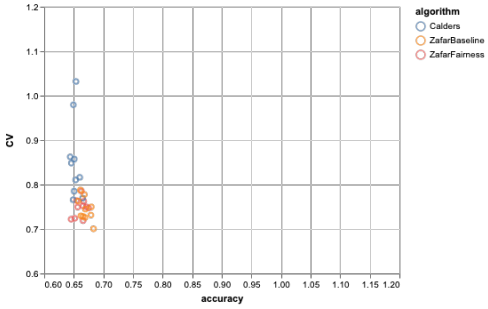



(c) Sex-Race

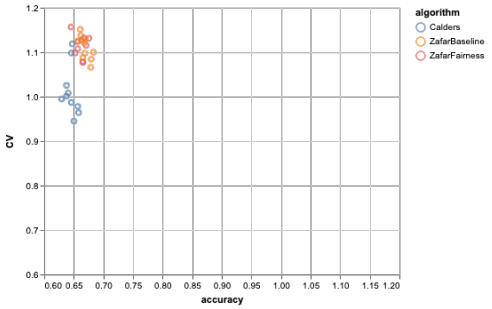
FIGURE 5. Conviction DIbinary vs Accuracy.




(a) Race

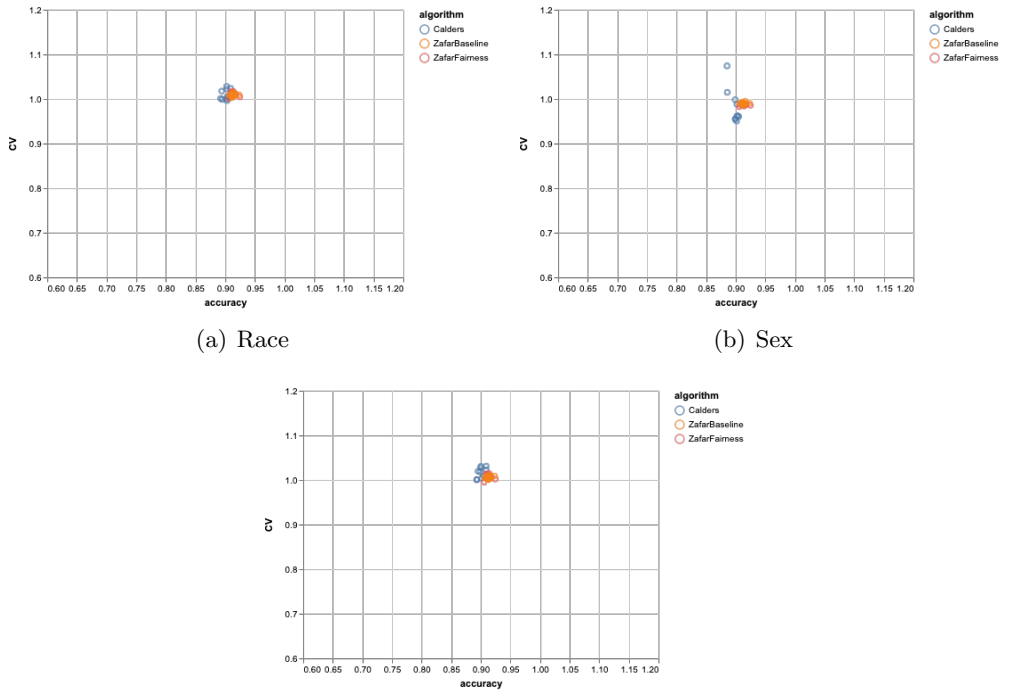



(b) Sex




(c) Sex-Race

FIGURE 6. Rearrest CV vs Accuracy.




(a) Race

(b) Sex

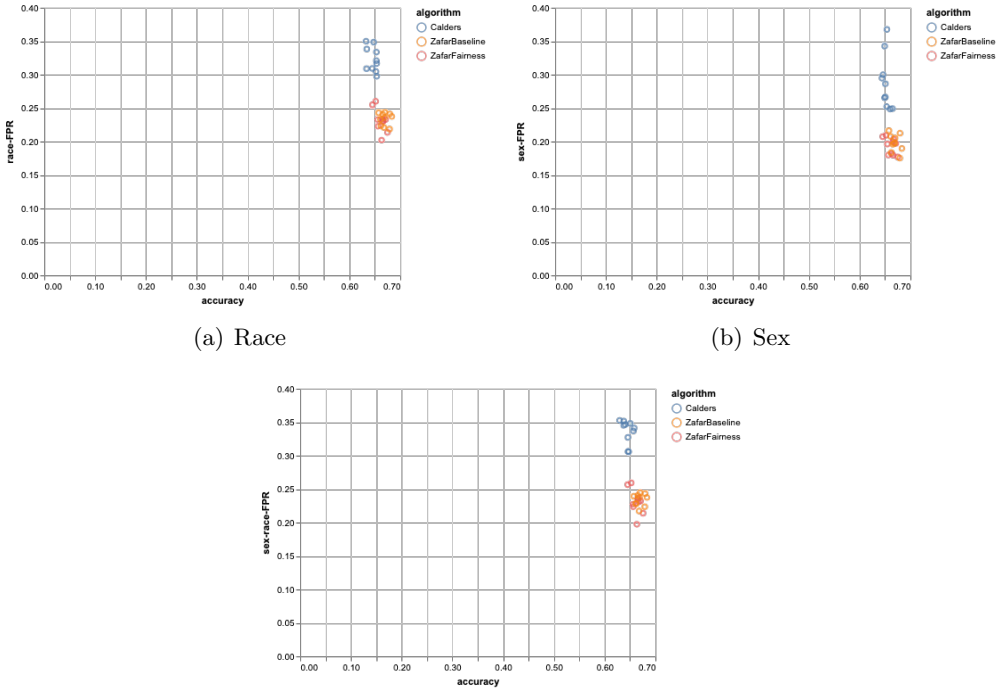
(c) Sex-Race

FIGURE 7. Conviction CV vs Accuracy.

The results for CV, seen in Figures 6 and 7, show that for race, sex and sex-race, re-arrest is pretty fair for both Calders' and Zafar's algorithms as they are all at around 1.0 which means that using conviction data makes the classifier very fair for both unprivileged and privileged defendants.

Comparing the two different fairness measures, DIbinary and CV, shows that CV is much fairer than looking at disparate impact. There is hardly any bias for race, sex and sex-race when looking at CV between re-arrest and convictions where convictions is the fairest. Meanwhile for DIbinary for both re-arrest and convictions, especially for convictions, the results show there is heavy bias towards one group for all race, sex and sex-race . This is most particular as Friedler et al.'s paper mentions that since Calders and Zafar algorithms are designed to optimize DI or CV, optimizing for one can be expected to optimize for the other, which is not what is seen here.

I then wanted to compare how the false positive rates and false negative rates where each point is the average of the positive and negative sensitive feature False Negative Rates and False Positive Rates from the 2/3 and 1/3 splits.



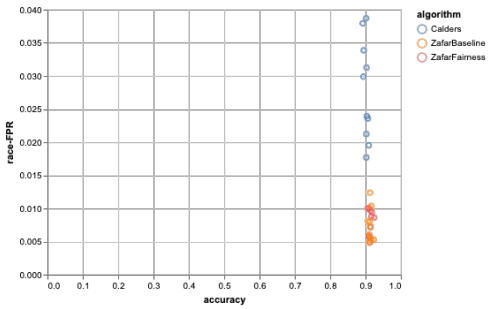

(a) Race

(b) Sex

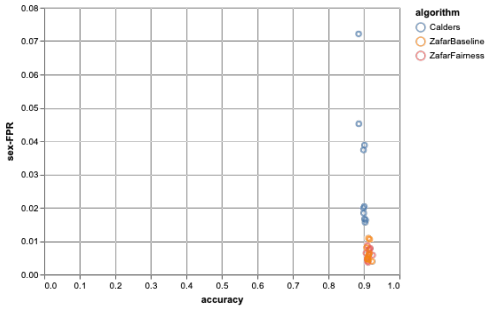
(c) Sex-Race

FIGURE 8. Rearrest FPR vs Accuracy. For all three metrics, Zafar et al.'s algorithms, both ZafarBaseline and ZafarFairness, have a lower FPR than Calders' algorithm.

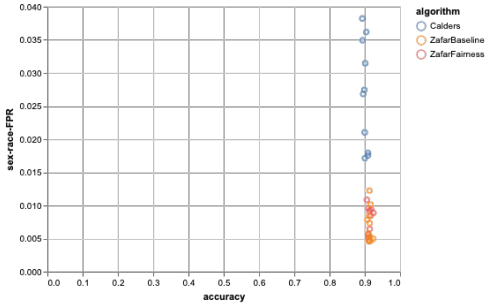
Where sex-FPR, race-FPR and sex-race-FPR, are the average of 0-FPR and 1-FPR for that specific metric. For example for sex-FPR, it takes the average of False Positive Rate of females, the negative sensitive attribute, and males, the positive sensitive attribute.




(a) Race

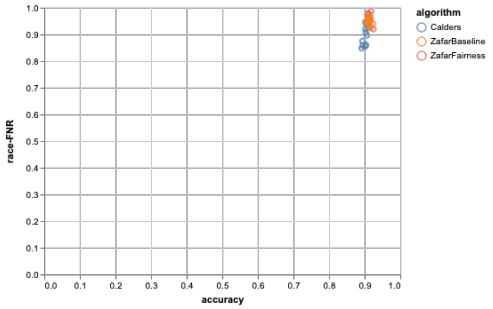



(b) Sex

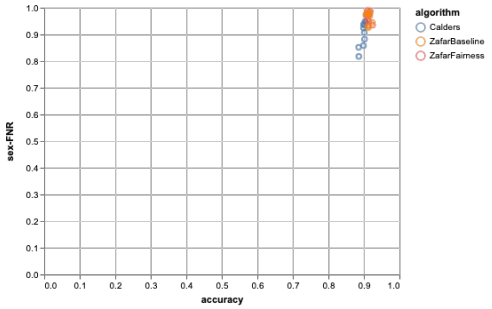



(c) Sex-Race

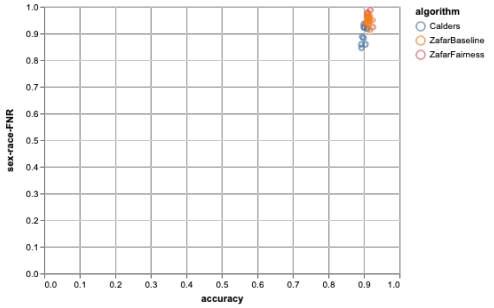
FIGURE 9. Conviction FPR vs Accuracy. For all three metrics, Zafar's algorithms have lower FPR than Calders' algorithm. It can also be seen that Calder's algorithm has no apparent pattern as most of the points are spread out especially for race and sex.




(a) Race

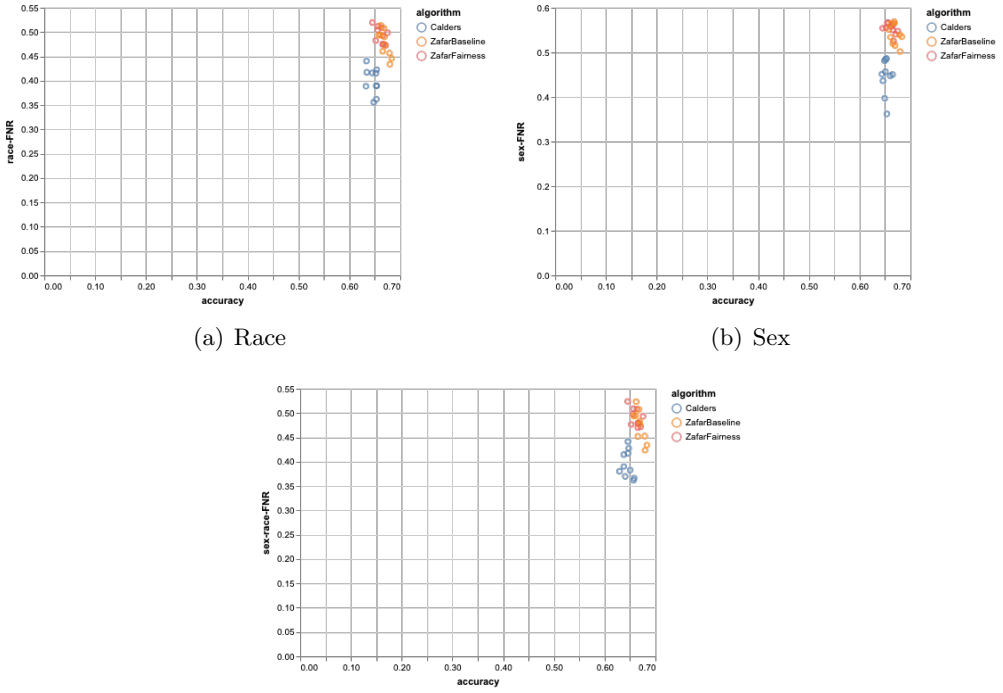



(b) Sex




(c) Sex-Race

FIGURE 11. Conviction FNR vs Accuracy. For all three metrics, Calder's algorithm has a lower FNR than Zafar's algorithm.




(a) Race

(b) Sex

(c) Sex-Race

FIGURE 10. Rearrest FNR vs Accuracy. For all three metrics, Calder's algorithm has a lower FNR than Zafar's algorithms however they are all between 0.35 and 0.55.

Where sex-FNR, race-FNR and sex-race-FNR, are the average of 0-FNR and 1-FNR for that specific metric. For example for sex-FNR, it takes the average of False Negative Rate of females, the negative sensitive attribute, and males, the positive sensitive attribute.

Figures 8 and 10 show the data for re-arrest False Positive Rate(FPR) and False Negative Rate(FNR) respectively while Figures 9 and 11 show the data for conviction FPR and FNR respectively. The Zafar et al.'s algorithms have a lower FNR and higher FPR than Calders for all data points in both re-arrest and convictions. Zafar et al.'s algorithms do better at optimizing for the false positive and false negative rates because these algorithms maximize fairness subject to false negative rate and false positive rate fairness constraints. This was done for ZafarFairness algorithm by solving the covariance function

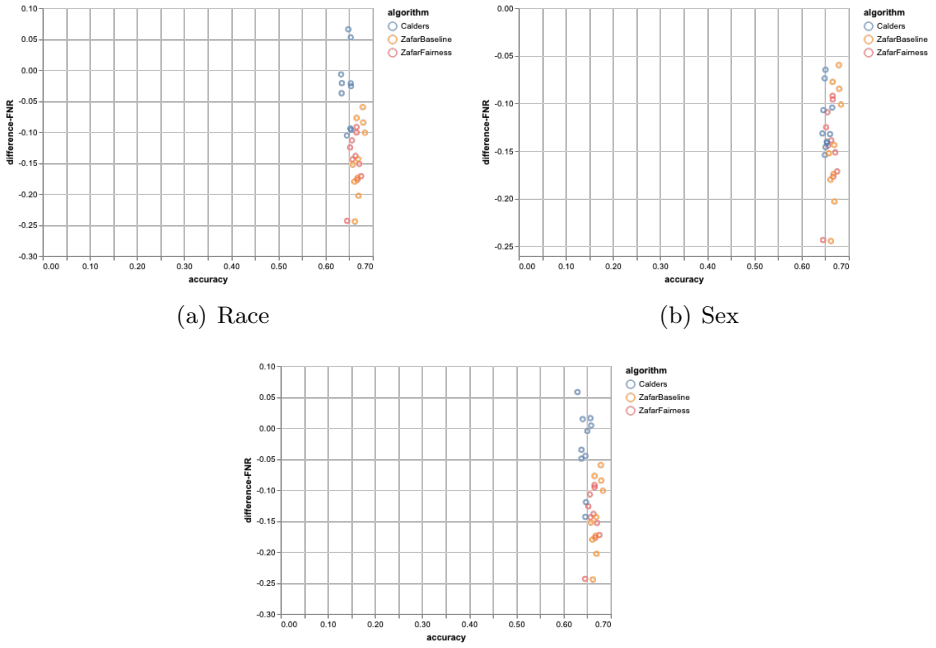
$$g_\theta(Y, x) = min(0, \frac{1-Y}{2} Y d_\theta(x))$$

for false negative rates and

$$g_\theta(Y, x) = min(0, \frac{1+Y}{2} Y d_\theta(x))$$

for false positive rates for each split of the training set. ZafarBaseline algorithm does this by training a classifier minimizing a loss function using a fair baseline decision boundary and then selects the set of misclassified points in the training data set having the sensitive attribute and binary class attribute set to what we are optimizing for, either for FNR or FPR. Meanwhile, Calders' algorithm tries to reduce the CV measure to as small as possible for each $R$ and does not try to optimize for FNR and FPR.

Going from re-arrest to convictions, the FPR or FNR drastically changes which is also seen in Jai's and mine's results. The FPR and FNR are most similar for re-arrest data which reflects what was seen in ProPublica's results and my results for race, sex, and sex-race, however, FPR and FNR for convictions data is opposite to what Jai and I have found. Jai and my results for race, sex and sex-race show that the average FPR is about 0.9 and close to 1 while the average FNR is very close to 0 and 0.1. However, for both algorithms FPR is closer to 0 and 0.1 while FNR is closer to 0.9 and 1.



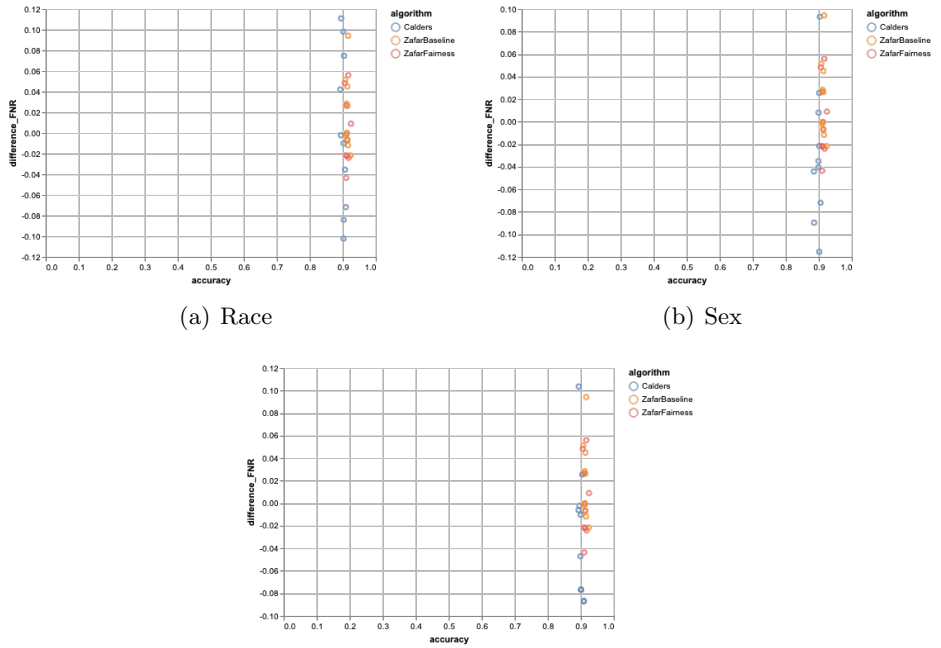

(a) Race

(b) Sex

(c) Sex-Race

FIGURE 12. Rearrest Difference FNR. For both race and sex-race, Zafar et al.s' algorithms have a lower $D_{FNR}$ than Calder's algorithm as it is the closest to 0. Meanwhile for sex, they both suffer from disparate mistreatment equally.

Zafar et al. discussed how a classifier does not suffer from disparate mistreatment if $D_{FPR}$ and $D_{FNR}$ is closer to 0, thus I took the definition of these two variables and plotted




(a) Race

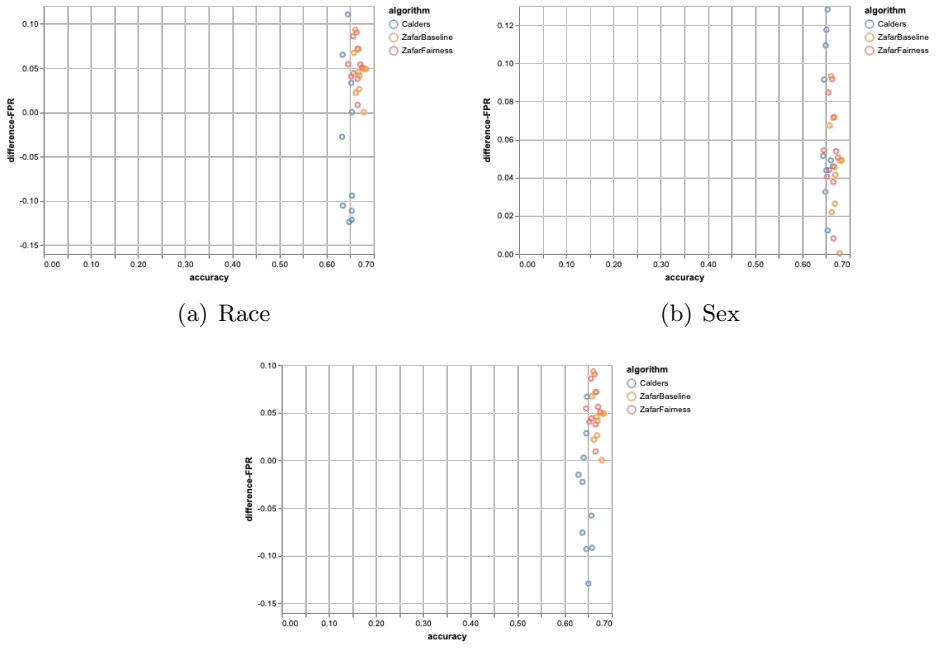
(b) Sex

(c) Sex-Race

FIGURE 13. Conviction Difference FNR. For all three metrics, all three algorithms range by the same scale thus for convictions there is more disparate mistreatment.

them which is shown in Figures 12 through 15. Looking at these figures it is obvious both datasets suffer from disparate mistreatment. However, going from re-arrest to convictions for $D_{FPR}$ and $D_{FNR}$, the range of $D_{FPR}$ and $D_{FNR}$ gets smaller and the points get closer to zero for both of Zafar et al.'s algorithms but not for Calder's algorithm as there is wider range. This shows that the convictions dataset suffers less from disparate mistreatment than ProPublica's dataset when looking at Zafar et al.'s algorithms. As stated before Zafar et al.'s algorithms actively try to optimize for FPR and FNR constraints while Calders does not, thus Zafar et al.'s algorithms do a better job at detecting disparate mistreatment as the range is smaller for their algorithms. For race, sex and sex-race, going from re-arrest to conviction dataset, the values for $D_{FPR}$ and $D_{FNR}$ get closer to 0 which shows that the conviction dataset does not suffer as much from disparate mistreatment as re-arrest dataset.

To make sure that this big jump of accuracy from rearrest to convictions is justified, I did some more analysis on other accuracy measures. As seen from table 4, the dataset demographics showed that there were about 5% of Whites and about 10% of African Americans convicted, thus about 90% of the defendants were not reoffending. This then




(a) Race

(b) Sex

(c) Sex-Race

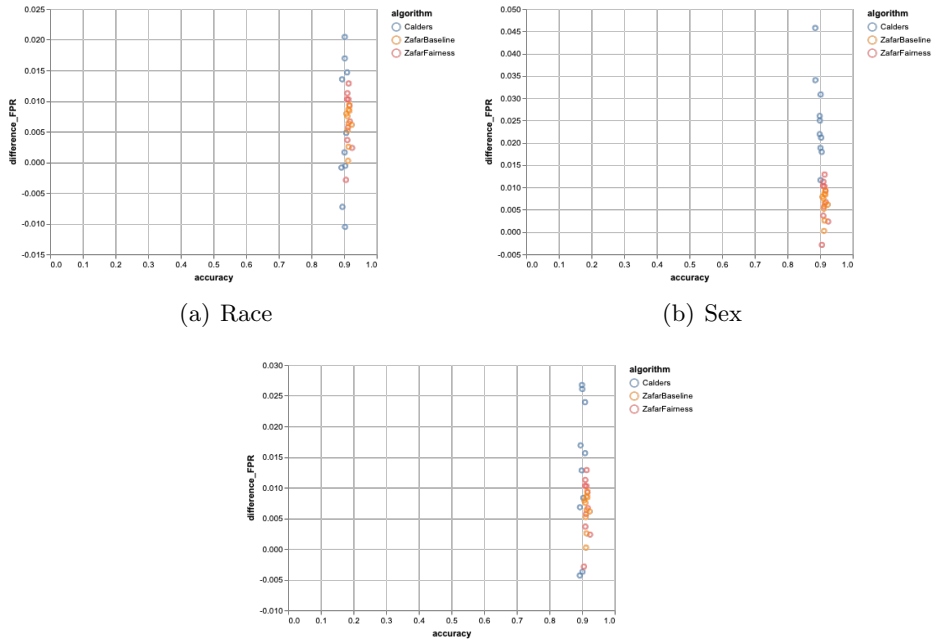
FIGURE 14. Rearrest Difference FPR. For race and sex-race, Zafar's algorithm has a lower $D_{FPR}$ then Calder's algorithm, which is closer to 0. Meanwhile for sex, they both suffer from disparate mistreatment equally.

could make the classifiers we analyzed classify all the defendants in our dataset as non-offending thus causing the accuracy of these classifiers using our dataset to be so high. This then could lead to the discrimination score, Calders' CV score, to be 1 for all the classifiers when using convictions data as well as if all defendants are non-offending then no discrimination as Calder's algorithm actively tries to have the discrimination score to be close to 0 as possible. To test this I looked at another measure of accuracy which is unweighted per class called balanced classification rate (BCR) defined as:

$$\frac{P[\widehat{Y} = 1|Y = 1] + P[\widehat{Y} = 0|Y = 0]}{2}$$

where $Y$ is the binary classification label where 1 is the privileged class and 0 is the negative class and $\widehat{Y}$ is the predicted outcomes of some algorithm.

Figure 16 shows the results of balanced classification rate using our convictions dataset, on the x-axis as BCR, vs ProPublica's rearrest dataset, on the y-axis as $rearrest\_BCR$. As shown it confirms our hypothesis that the Calders' and Zafar et al.'s classifiers are classifying everyone as not reoffending and thus not as accurate as we thought as the




(a) Race

(b) Sex

(c) Sex-Race

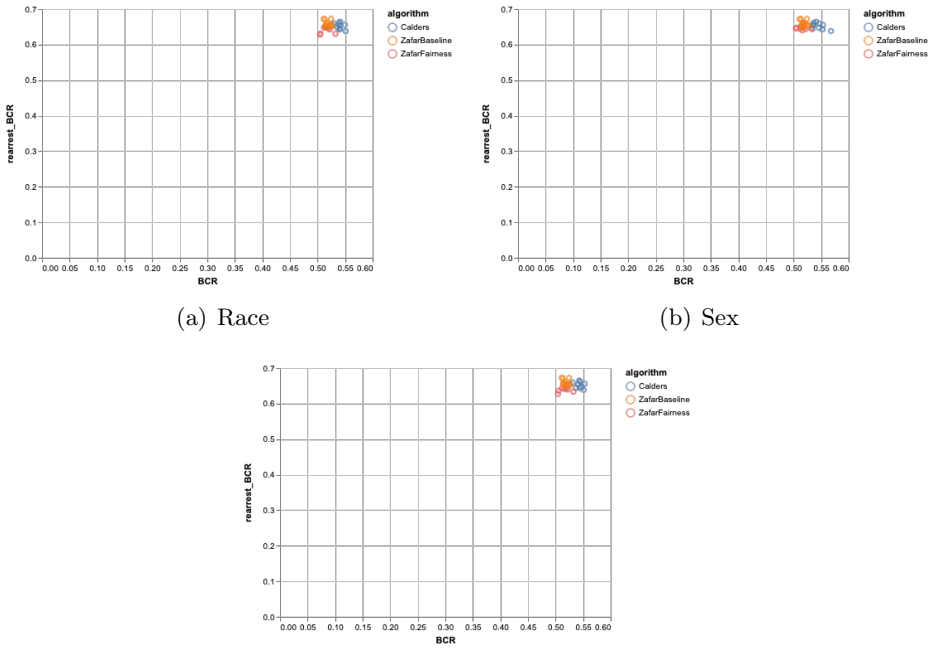
FIGURE 15. Conviction Difference FPR, For all three metrics the range is on the same scale thus there is more disparate mistreatment for convictions.

BCR scores, for race, sex and sex-race, using our dataset is lower than BCR scores using ProPublica's rearrest dataset.

## 5. Conclusion and Future Work

As seen looking at convictions rather than re-arrest for the intersectionality of sex and race failed to yield significantly different results, especially for false positive rates. Since the base rates for sex and race were not the same, especially for Males, it was predicted that there is still biased towards African-Americans, male and female. This may due to the fact, as Jai pointed out in his paper, that many of these defendants agree to plea deals thus would not be convicted of a crime which would confound the data especially since I did not have the data of if the defendant was convicted of the crime they committed of the original COMPAS screening.

However, looking at convictions rather than re-arrest for different classifiers did yield significant results. Using conviction rather than re-arrest allowed the classifiers to be significantly more accurate, however, this was shown that it was because the algorithms classified all defendants as non re-offending. By looking at other types of accuracy measures we saw that using conviction data instead of rearrest data the dataset is still biased towards one group and still has discrimination. This could be solved by being able to get the data




(a) Race

(b) Sex

(c) Sex-Race

FIGURE 16. Balanced Classification Rate, For all three metrics, both algorithms range from 0.5 to 0.55 for our convictions dataset, which is labeled under BCR, and range from 0.6 and 0.7 for ProPublica's rearrest dataset, which is labeled under $rearrest\_BCR$.

from Broward County for each defendant in ProPublica's dataset in order to see if each case actually went to court or not. It was seen that looking at convictions caused more disparate impact while also causing there to be little to no discrimination. There was still disparate mistreatment when looking at convictions, however, there was less than when looking at re-arrest data. In the future it would be interesting to look at other classifiers such as Hardt et al.'s algorithm in order to compare it with Zafar et al.'s classifiers in order to see if Zafar et al.'s results changed, however, I hypothesize that Hardt et al.'s classifier will still do better especially since our dataset is smaller than ProPublica's dataset.

## REFERENCES

[1] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21:277–292, 2010.

[2] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

[3] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[4] Toon Calders Faisal Kamiran and Mykola Pechenizkiy. Discrimination aware decision tree learning. *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 869–874, 2010.
[5] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *In Advances in neural information processing systems*, pages 3315–3323, 2016.
[6] Lauren Kirchner Jeff Larson, Surya Mattu and Julia Angwin. How we analyzed the compas recidivism algorithm propublica, May 2016.
[7] Surya Mattu Julia Angwin, Jeff Larson and ProPublica Lauren Kirchner. Machine bias, 2016.
[8] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
[9] Manuel Gomez Rodriguez Muhammad Bilal Zafar, Isabel Valera. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web - WWW 17*, 2017.
[10] Jai Nimgaonkar. Re-evaluation of the propublica article on machine bias. page 28, Decemeber 2018.
[11] Suresh Venkatasubramanian Sonam Choudhary Evan P. Hamilton Sorelle A. Friedler, Carlos Scheidegger and Derek Roth. fairness 0.1.8.
[12] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.