Re-Exploring COMPAS Bias

Zekai Fan

Violet Yao

May 2019

1 Introduction

In the beginning of the semester, we read an article by ProPublica about the recidivism-predicting algorithm COMPAS, or Correctional Offender Management Profiling for Alternative Sanctions, developed by the private company Northpointe, known now as Equivant. The article entitled "Machine Bias" [2] claims that COMPAS scoring algorithm is biased against African Americans, as it is particularly likely to falsely flag black defendants as future criminals, almost at twice the rate as white defendants. For example, arrested both for drug possession, Fudget, a white defendant, receives a score of 3, which belongs to low-risk category; while Parker, a black defendant, receives a score of 10, which belongs to high-risk category.

We believe that algorithm-based tools, such as COMPAS, help judges

decide. Humans are prone to mistakes when tired or distracted, while algorithm just iterates through what it is programmed to do and thus incapable of neglecting anything that it is supposed to do. Moreover, human judges may be unknowingly biased against a certain group or groups because of their idiosyncratic background or prior experiences. Judge A, from background A', may make a decision A" over subject S, while Judge B, from background B', may make a decision B" over the same subject S. Furthermore, in a less open and democratic environment, a judge's decision may be affected by politics or personal interest. Algorithms, on the contrary, are more consistent in decision making and can be made more transparent than human judges. Thus, we do believe in Data's power to transform law practice and advance justice.

In response to ProPublica, Brennan, founder of COMPAS, states that COMPAS does not use race as a factor. Instead, COMPAS score is derived from a large set of questions such as "How often did you get in fights while at school?", "Do you agree or disagree with the statement that a hungry person has a right to steal", etc. It brings out our research question: if we assume that developers of COMPAS are not inherently biased against African Americans, why do we see the highly biased results given by ProPublica's analysis?

At first, we want to dive into the criteria used by the algorithm to make a decision. While we know that demographic characteristics are not a part of the set of decision variables, we do not know if there are any decision variables that are strongly correlated with demographic characteristics. However, we find that this information has never been disclosed. Northpointe is a for-

profit company and considers the algorithm proprietary. Thus, instead of studying the algorithm itself, we have to study the result of algorithm - COMPAS decile scores.

Then, where can we get COMPAS scores? Again, Northpointe does not disclose any data, including COMPAS scores and recipient profiles. But we gain access to COMPAS scores for Broward County in Florida from 2013 to 2014. This dataset [4] is made accessible by ProPublica, who made an open-record request to Broward County.

2 Exploratory Analysis

In class, we learned to use the bias and fairness audit toolkit "Aequitas" [5], developed by the Center for Data Science and Public Policy at the University of Chicago. The developers of Aequitas used the same Broward County dataset used by ProPublica as a demonstration for the usage of their software. Aequitas is able to display various rate disparities for different groups in a dataset. According to the decision tree proposed by its developers, we should select "false positive rate" if we are interested in "fairness in errors" for "punitive interventions" that affect more than "a small percentage of population". We concur with this result, as we believe if false positivity is inevitable, every group of people should be equally susceptible to it. The acceptable 80% to 125% interval is marked by the red lines (Figure 1). It should be noted that some experts have a different opinion. The developers of COMPAS selected false discovery rate as a benchmark to calibrate their

algorithm on. We reproduced the analysis of false positive rate on the original Broward County dataset, using Aequitas.

2.1 Sex

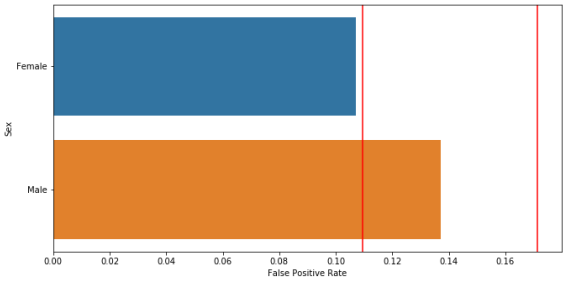
We select the category "Male" as the baseline. It has a false positive rate of 13.7%. "Female" has a 10.7% false positive rate, lower than 80% of that of "Male". This indicates that all groups fail the audit. (Figure 1a)

2.2 Age

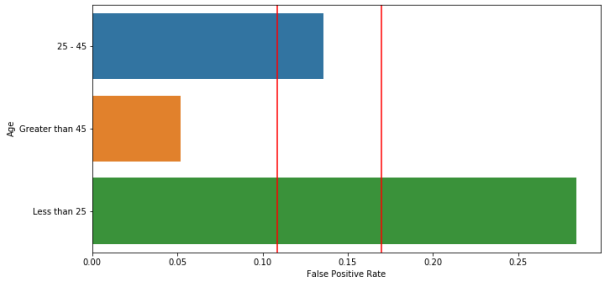
We select the category "Greater than 45" as the baseline. It has a false positive rate of 13.5%. "Greater than 45" has a 5.1% false positive rate, far lower than 80% of that of "25 - 45". "Less than 25" reports 28.4% false positive rate, far higher than 125% of that of "25 - 45". Again, all groups fail the audit. (Figure 1b)

2.3 Race

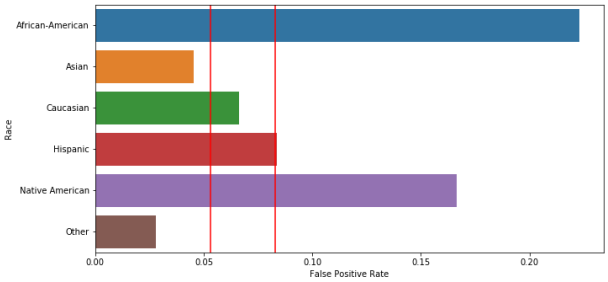
We select the category "Caucasian" as the baseline. It has a false positive rate of 6.7%. "African-American" has a 22.3% false positive rate, far higher than 125% of that of "Caucasian". "Asian" has a false positive rate of 4.5%, lower than 80% of that of "Caucasian". "Native American" has a false positive rate of 8.3%, far higher than 125% of that of "Caucasian." "Hispanic" is the



(a) Sex



(b) Age Category



(c) Race

Figure 1: False Positive Rates for Different Groups

only group that falls in the range of 80% - 125% of baseline determined by "Caucasian". For yet another time, all groups fail the audit. (Figure 1c)

3 Dataset Visualization

It is worth noticing that not all Broward County residents have COMPAS score. Only those who were arrested would be assigned COMPAS scores. We are interested in how it compares to demographics in Broward County in general. We visit American FactFinder [1] for surveys about Broward County demographics. We find a significant mismatch between those who were arrested and received a COMPAS score and the general Broward demographics. At this stage, we were thinking that other demographic factors, such as age and gender, might also result in biased COMPAS scores.

3.1 Sex

While female-male ratio is roughly 1 to 1 in Broward demographics, the males make up three fourths in ProPublica's dataset. (Figure 2)

3.2 Age

While people aged 25 - 45 is a smaller portion in Broward Demographics, it makes up a much larger portion in ProPublica's dataset. Percentage of both

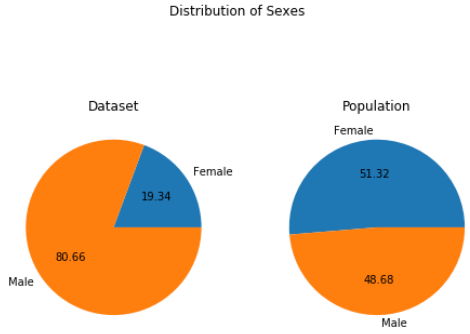


Figure 2: Comparison of Distributions of Sexes

"Greater than 45" and "Less than 25" in ProPublica's dataset are less than their corresponding representation in Broward Demographics. (Figure 3)

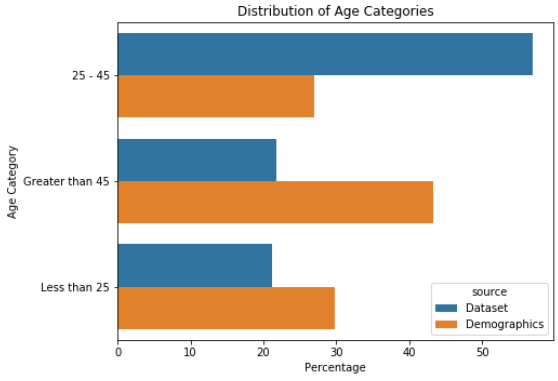


Figure 3: Comparison of Distributions of Age Categories

3.3 Race

The Broward County dataset considers "Hispanic" as a race. U.S. Census on the other hand considers "Hispanic" as a characteristic that people possess in addition to their races. In order to make them comparable, we transform the data from U.S. Census so that they align with those on the Broward County dataset. In this process, we assume that people belonging to only one race are equally likely to be Hispanic. We categorize people who belong to two or more races as "Other".

While African American are 21% of Broward Population, it makes up over 50% of ProPublica's dataset. In contrast, compared to their demographic representation, Caucasian, Hispanic and Asian have a lower percentage in ProPublica's dataset. (Figure 4)

4 Analysis of Explanatory Power

In order to show the extent to which COMPAS scores are associated with demographic characteristics, we first establish a baseline: to what extent recidivism is associated with demographic characteristics. To do so, we run a linear regression in which demographic characteristics are independent variables and two-year recidivism dependent variable. The formula is $y_{recid} = \beta_1 X_{sex} + \beta_2 X_{age} + \beta_3 X_{race} + \epsilon$, where X_{sex}, X_{age}, and X_{race} are one-hot encodings. We fit the model on the Broward County dataset, and found a R^2 score of 0.374 (Figure 5, red line). This implies that without any predictive

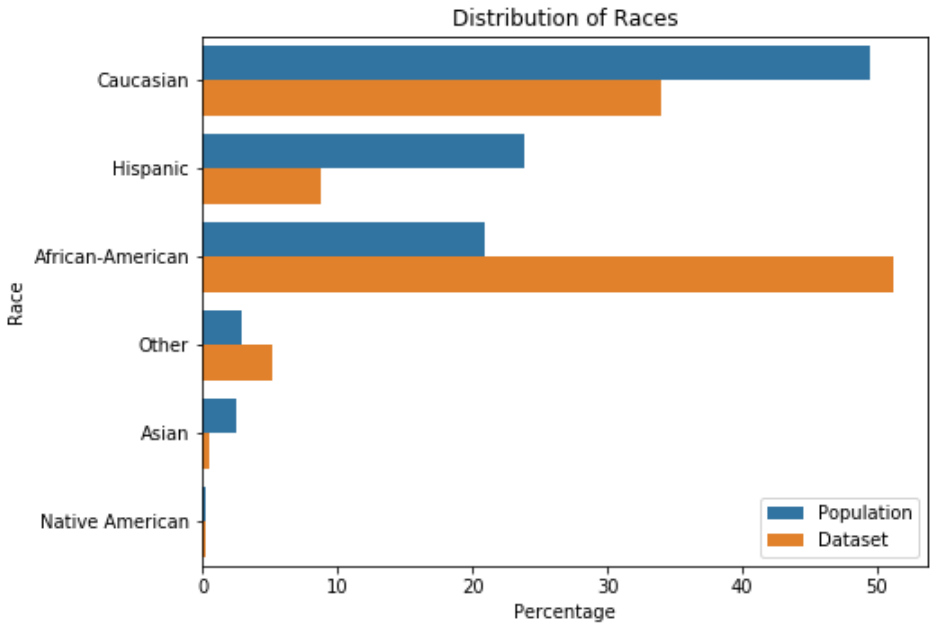


Figure 4: Comparison of Distributions of Races

model, recidivism to a moderate extent can be explained by demographic characteristics. This effect can be attributed to unbalanced dataset as well as various degree of propensity to recidivate due to difference in different groups of people, though the reason behind it is out of the scope of this project. For example, higher level of testosterone in men may be associated with more aggression; underdeveloped prefrontal cortexes in people age less than 25 may be associated with more impulsive behaviors.

We then take into consideration COMPAS scores. To do so, we condition on COMPAS score and rerun the regression. For each of the ten possible

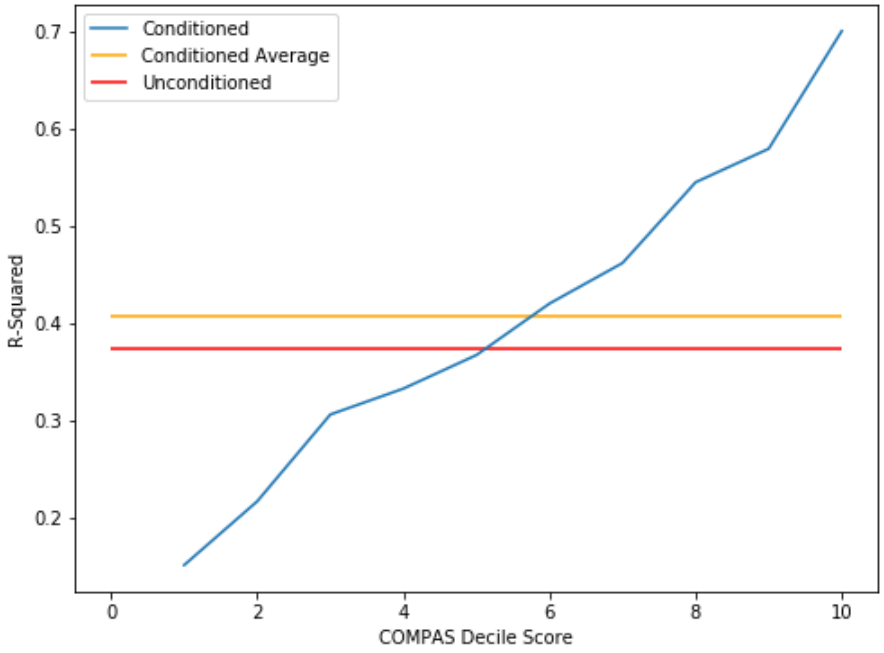


Figure 5: R^2 Scores of All Demographic Characteristics Considered

scores, we fit the model on rows with that score only, so we obtain ten R^2 scores (blue line), one for each COMPAS score. The average of the ten R^2 scores is 0.408 (orange line), higher, but not much, than the unconditioned model's 0.374. This implies that on average, COMPAS does not make demographic characteristics more powerful in explaining recidivism than they already are by too much.

The R^2 scores, however, have a high variance. The lowest R^2 score corresponds to COMPAS score 1. It is only 0.153, not as much as half of the average R^2 score of 0.408. The highest R^2 score, 0.700, corresponds to

COMPAS score 10. The higher a COMPAS score is, the higher the R^2 score. This implies that while on average the algorithm does not make demographic characteristics more explanatory by much, the higher a score it gives, the more explanatory power we observe from the demographic characteristics.

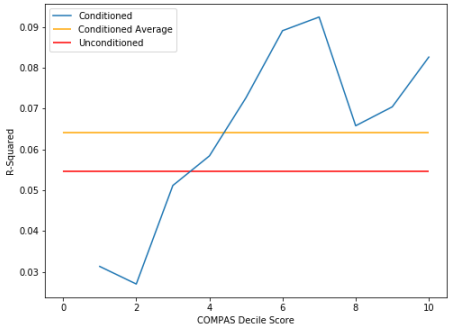
This effect makes us see the groups with higher recidivism risk in general more likely to have their recidivism explained by the demographic characteristics that define the groups.

For example, if "fewer years of education" lead to a higher COMPAS score, people age less than 25, who have the highest rate of observed two-year recidivism (56.5% compared to 46.0% people age 25 - 45's and people age greater than 45's 31.6%), have their recidivism explained by being younger than 25 better with the algorithm than without, because fewer years of education are associated with being younger than 25. Since people age below 25 have a disproportionately high representation in the Broward County dataset, we observe a high bias towards them, even if "the number of years in education" by itself is a good predictor of recidivism.

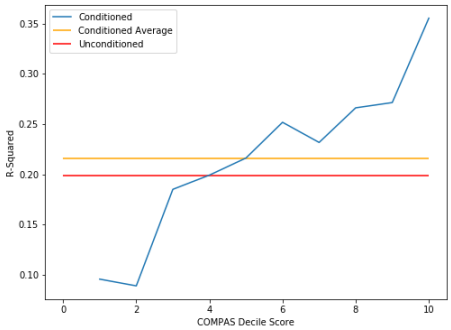
We repeat the process for each of the three demographic characteristics independently and observe similar trends. (Figure 6)

5 Analysis of Variability

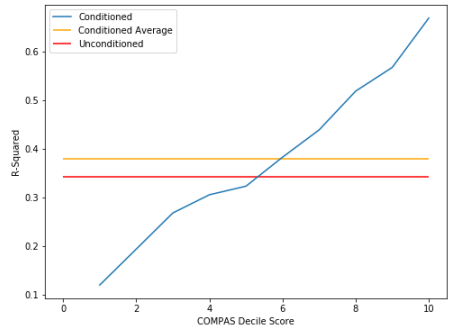
To find out the reason behind the positive correlation between R^2 scores and COMPAS scores, we become interested in the algorithm's performance



(a) Sex



(b) Age Category



(c) Race

Figure 6: R^2 Scores for Different Groups

across its scores. To do so, we formulate COMPAS scores as probability to recidivate. A score of 1 represents 10% probability of recidivism, and a score of 10 dictates with certainty that the individual will recidivate in two years. Note that under our formulation, a score of 5 predicts with equal probability that the individual recidivates or not.

To measure the algorithm's performance, we select Brier score as the measurement for model loss. A Brier score is defined as $BS = \frac{1}{N}\sum_{i=1}^{N}(pred_i - obs_i)^2$. In our case, there are two possible outcomes, recidivism and no recidivism. Brier scores here can be thus be simplified to $BS = (recid - COMPAS)^2$. A score of 0 indicates perfect prediction, while a score of 1 is the worst possible. A score of 0.25 (Figure 7, red line) is the same as a coin flip. We calculate a Brier score for the algorithm on the entire Broward County dataset, and then we calculate an average Brier score for each COMPAS score (blue line), aggregating all rows with that COMPAS score. Finally, we calculate an average score for the ten Brier scores (orange line).

The 0.229 average score indicates that the algorithm does a better job predicting two-year recidivism than flipping a coin by a moderate margin. However, there is ample room for improvement. More importantly, we observe heteroskedasticity in COMPAS scores in the Broward County dataset. In general, the algorithm's predictivity gets worse as it gives a higher score. The 0.181 Brier score for COMPAS score of 1 indicates very good model predictivity for low risk individuals, while the algorithm's prediction is worse than a coin flip when it yields a score of 7 ($BS = 0.254$) or 9 ($BS = 0.251$). This implies that the algorithm's performance gets worse as the risk of re-

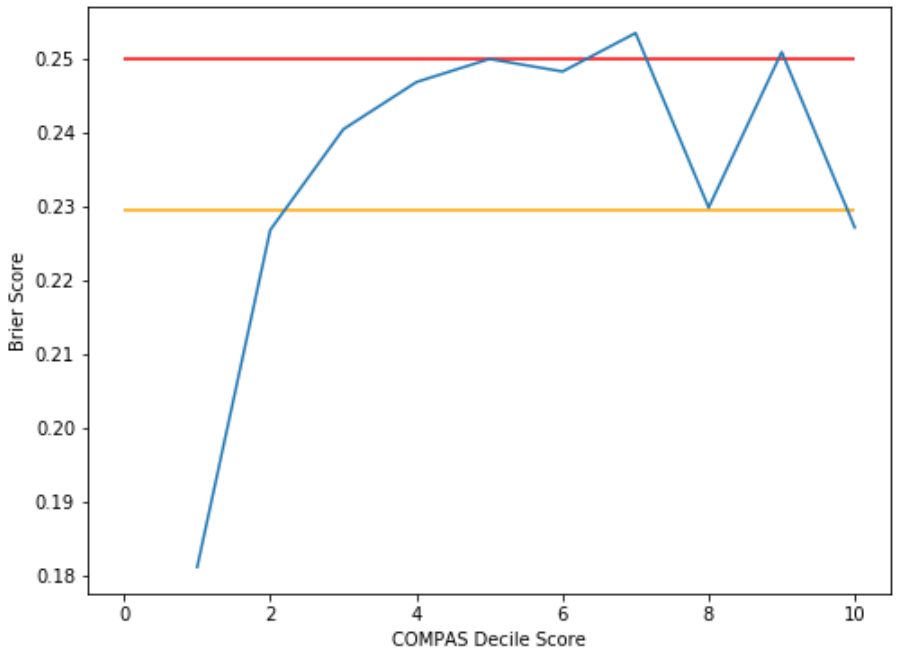


Figure 7: Heteroskedasticity of COMPAS Scores

cidivism gets higher, in general.

It could explain the observed false positive rate disparities across demographic groups, in that the algorithm predicts more accurately two-year recidivism for groups with lower average risk and less accurately for those with higher average risk. That is, although the causation behind this effect may be complicated, it could be the case that the algorithm is not necessarily biased against some specific demographic groups by themselves, but it is biased against groups with higher risk instead. Under this assumption, one way to reduce the algorithm's bias on specific groups is to improve its

productivity for high risk individuals, in a way such that it becomes more conservative when giving a high risk score to an individual.

6 Bootstrapping

As an effort to remedy the unbalanced Broward County dataset, we decide to bootstrap. Due to the limitations in the dataset, especially its lack of high-risk observations for many demographic groups, we have to choose between including more samples in each iteration and including more groups.

We settle to the following procedure: randomly draw 200 samples from each set in a demographic category, 100 of which are labeled high risk, 100 low risk. For example, for the demographic characteristic "sex", we randomly select 100 low-risk males, 100 high-risk males, 100 low-risk females, and 100 high-risk females. Note that the sampler is outcome-agnostic (i.e. it does not consider if the individual recidivates or not when deciding whether or not to include them in the bootstrapped sample). We then compute false positive rate for each group on the sample. A few examples of samples are shown in Figure 8. The procedure is done 500 iterations before an average false positive rate is calculated. We consider the bootstrapped mean false positive rate to be representative of a balanced dataset.

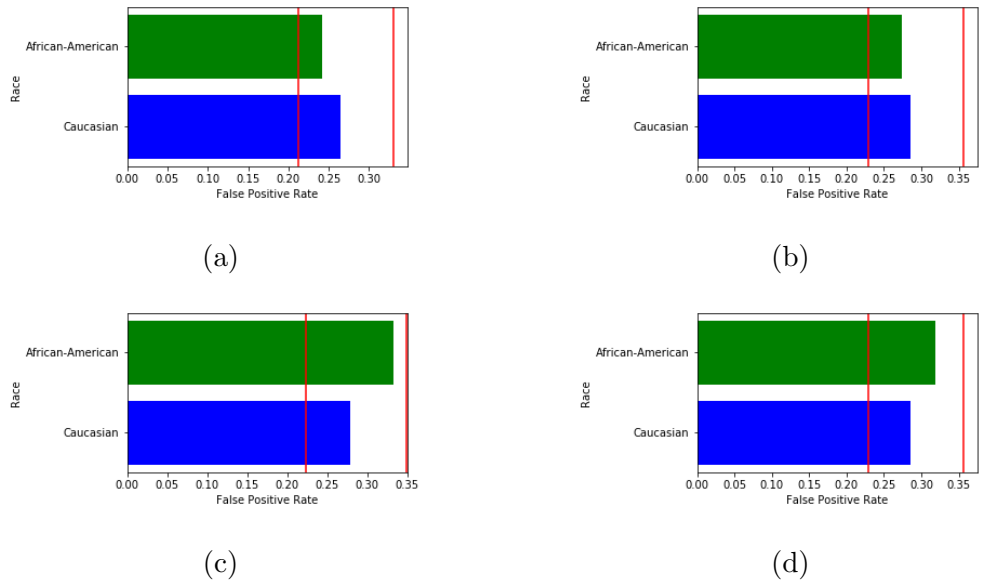


Figure 8: Four Samples

6.1 Sex

We select the category "Male" as the baseline. It has a false positive rate of 28.5%. "Female" has a 34% false positive rate, which falls within the acceptable interval of 80% - 125% of that of "Male". (Figure 9)

6.2 Age

We select the category "25 - 45" as the baseline. It has a false positive rate of 29%. "Greater than 45" has a 33% false positive rate and "Less than 25" reports 32% false positive rate. Both fall into the acceptable interval of 80% - 125% of that of the reference group. (Figure 10)

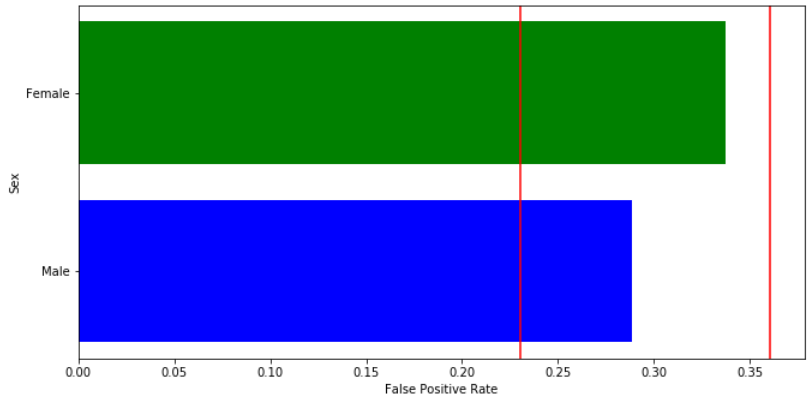


Figure 9: Average FPR Computed on Bootstrapped Samples for Sex

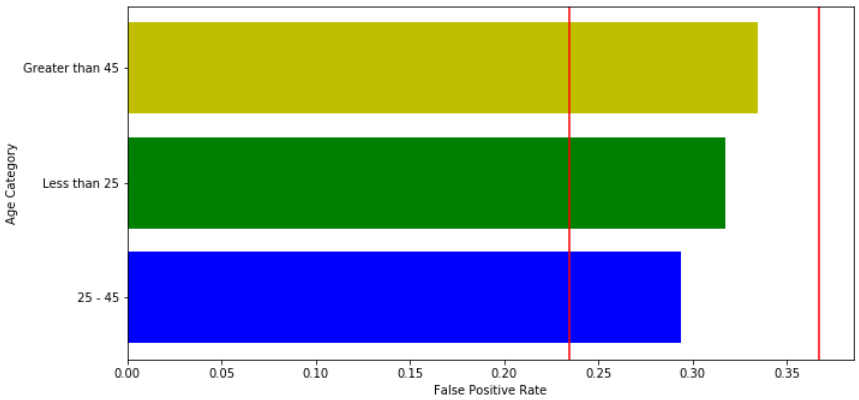


Figure 10: Average FPR Computed on Bootstrapped Samples for Age

6.3 Race

We have to drop all groups other than "Caucasian" and "African American" because there are fewer than 100 observations for high-risk for any of the other

groups in the dataset. We select the category "Caucasian" as the baseline. It has a false positive rate of 29%. "African American" has a false positive only slightly higher than that of "Caucasian", which again falls within the acceptable interval. (Figure 11)

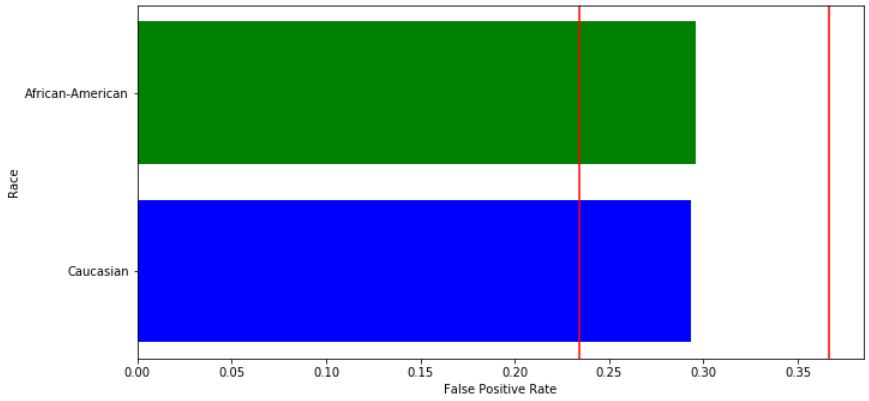


Figure 11: Average FPR Computed on Bootstrapped Samples for Race

7 Ad Absurdum: An Example

It seems clear now that blindly applying an audit tool may not result in sound conclusions about bias in models, as bias is likely to reside in an unbalanced dataset. Even though a dataset is not intentionally biased, as a convenience sample (e.g. Broward Country, Florida), it does not faithfully represent the population (e.g. United States). As a result, many assumptions we have for the population may not hold for the convenience sample.

In the Broward County dataset, each individual's birth date is recorded. We infer each individual's zodiac sign (e.g. Ares) from their birth date and attempt to run Aequitas for that label. Although zodiac signs are believed by some people historically or for entertainment purposes to be associated with personality, it is well perceived as a pseudoscience. We assume that zodiac signs are independent from any and all characteristics of an individual, including their criminal propensity. However, Aequitas shows us a different story.

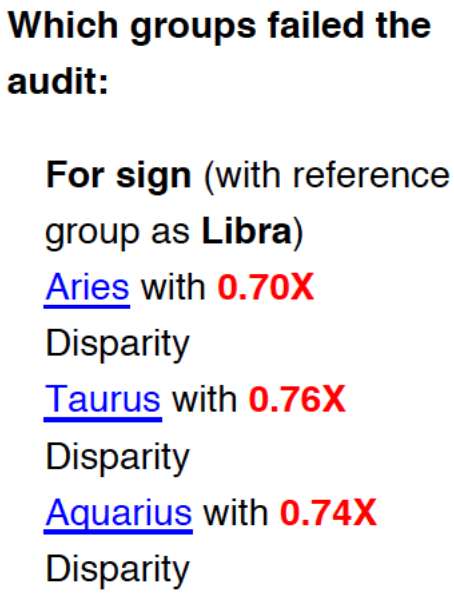


Figure 12: Aequitas Report According to Astrology

Libra being reference, Aries, Taurus, and Aquarius fail the audit for false positive rate parity. Unless one actually believes in the "ram-like" amity an Aries person possesses, this audit should mean nothing more than an unbalanced dataset and noises in applying algorithms. We would like to use

the absurdity of this example to show the importance of a balanced dataset using automated tools to determine model bias. In addition, one may also see that a seemingly balanced dataset may be biased in some aspects. We must closely examine a dataset before imposing our usual assumptions.

8 Conclusion

We find ProPublica's claim about racial bias in COMPAS can be largely attributed to the bias of the Broward County dataset that they used for analysis, instead of COMPAS itself. The dataset violates our assumption for a balanced dataset in many ways, including but not limited to: 1. mismatch between the distribution of demographic characteristics in the dataset and that of those in the Broward County population, and 2. difference in characteristics between the Broward County population and the population of the United States.

On the other hand, we also find COMPAS is biased against people with higher risk of recidivism. The algorithm exaggerates the risk of such individuals to an extent that is unrealistically high. As a result, we observe higher false positive rates in some groups, as those groups generally have a higher risk. It is not clear based on our analysis the cause of this bias. Nevertheless, we deem these as potential causes: 1. lack of observations of high-risk individuals in developing the algorithm, 2. loss of generality when decision variables that work well in predicting recidivism for people of lower risk are ported to predicting recidivism for people of higher risk, and 3. imprecise

definition of the scores themselves.

It is worth noting that there may exist trade-off between fairness and effectiveness. Strong predictors of recidivism may be highly correlated with certain categories of the demographics, even in very good models. If those predictors are included in a recidivism-predicting model, different groups of people will have different distribution of predicted levels of risk. If those strong predictors are excluded from the model, the model's accuracy in general may be adversely affected, while the its accuracies for different groups may converge, however to a lower value compared to if the strong predictors are included; its false positive rates for different groups may converge to a higher value than otherwise.

This effect lands decision makers to a hard choice between convergence and absolute efficiency. It can be remedied by a holistic review of the individual in question by the "end-users" of the software, the judges. If they elect to take into consideration a risk score, they should be informed of the criteria where the score is decided. They should decide the extent to which the general recidivism-predicting model applies to the individual in front of them, before letting their judgment, which has profound consequences on both the individual and the society, be largely influenced by a numerical value. For instance, if "highest degree attained" is a decision variable within a linear model, the individual's opportunity to education must be considered, as an individual from a wealthy family who only graduates community college may pose a higher, instead of lower, risk of recidivism, compared to one from a poor family who successfully graduates high school.

Although more complicated and non-parametric models may relieve the need for holistic review, the fact of them being more complicated will make them more susceptible to bias. Everyone is unique to some extent, so judges as a human factor are desirable in addition to faceless decision machines. The higher variance due to our humanity may well be a lesser evil.

9 Future Work

At present, we, as well as other researchers, are only able to study the result of COMPAS algorithm, instead of studying the algorithm itself. What's more, data sources are extremely limited. We only have COMPAS scores for Broward County 2013 - 2014 collected by ProPublica. We tried to look for more data to confirm our conclusion, but there are no other data sources in public domain. It is difficult to request data from COMPAS directly since it is a for-profit company and considers its algorithm proprietary. However, requesting data from the government should be viable. We sincerely hope that researchers, who are passionate about this topic and have access to the government database, could test our conclusion against more data.

We believe that COMPAS's data and algorithm should be accessible to the general public, just as the laws of a nation are publicly accessible and clearly articulated. A New York Times story [3] reported in 2017 that Mr. Loomis, who was sent to prison partly because of his COMPAS score, appealed that his right to due process was violated. Mr. Loomis's argument is strong. Although based on our research we concluded that COMPAS is not

biased against any demographic group, we cannot argue against Mr. Loomis that his sentence was fair. After all, we did not know anything more about COMPAS algorithm than Mr. Loomis.

We call for more government regulation and public discussion over Computational law. COMPAS scores are used in sentencing and thus can be considered as part of law. Law should be a consensus of citizens, right? What's more, shouldn't law be made by legislature? Is it ethical for a company to "commercialize" law?

References

[1] United States Census Bureau. American FactFinder.

[2] Angwin et al. Machine Bias. 2016.

[3] Adam Liptak. Sent to prison by a software program's secret algorithms. 2017.

[4] ProPublica. Data and analysis for 'Machine Bias', 2017.

[5] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.