Senior Thesis in Mathematics

Racism without a Face: Predictive Statistics in the Criminal Justice System

Author:

Luis Antonio Espino

Advisor:

Dr. Johanna Hardin

Submitted to Pomona College in Partial Fulfillment

of the Degree of Bachelor of Arts

April 30, 2018

Abstract

The use of statistical risk assessment tools in the criminal justice system is an increasingly pressing issue. From determining needs and allocating resources post-release, to helping determine how much time a person will spend in cages, wide-spread adoption of these predictive tools continues. Their expanded use and the promise of a more equitable and effective criminal justice system has garnered hope and excitement for some, while their real-world implementation is a concern for others. Recent studies have demonstrated that a prominent risk assessment tool, called COMPAS, is having disparate impact across different demographic variables. My thesis is motivated by two central questions concerning COMPAS and risk assessment tools more generally. First, what drives disparate impact across race? More specifically, what affect does biased data have on this disparate impact? I conclude that biased data, where bias is defined as over-representation of a group, leads to biased algorithms and therefore biased results in the form of disparate impact. Furthermore, I claim that these results indicate a need for analyzing and understanding the prison industrial complex and the interlocking systems of domination that lead to over-representation of black defendants in the criminal justice system. Moreover, these results beg for a multidisciplinary, comprehensive approach to both the building and implementation of risk assessment tools.

Acknowledgements

Much of what I do is informed and held up by my lived experiences as a first-generation, low-income Latino. In carrying out this thesis, I was able to do just that. I am grateful to Professor Hardin for bringing me into this project, for introducing me to the world of data science at the intersection of social justice, and for her continued support, advice, and patience in this process. Thank you for your support throughout these past four years. I also want to thank Professor Shahriari, whose advice and support were critical in my decision to major in math. A big thank you also to Professor Ochoa, for all your mentorship and encouragement in imagining and working towards a better world.

I am grateful to Vanessa Machuca, who has been a pillar of support for me, both in and out of the classroom. Thank you for all the laughs, late nights, and meaningful conversations that have helped sustain me along the way. Special shout out to my parents, who are the reason I got to study for a living these past four years. Especialmente a mi mama. Gracias por todo Jefiíta.

Contents

Chapter 1

Introduction

1.1 Exposing Systematic Injustice in the Criminal Justice System: COMPAS

On May 23rd, 2016, ProPublica, a Pulitzer Prize-winning newsroom of investigative journalism, launched an article that shed public spotlight on racial bias in a statistical risk assessment tool for recidivism known as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) [1]. Using data from over 10,000 people arrested in Broward County, Florida, Angwin et al. [1] demonstrated that while the algorithm's predictive accuracy was about the same for both black and white defendants, around 61%, the algorithm's failure affected the groups in different and opposite ways. The false positive rates for black defendants were about twice as high than for white defendants, 44.9% compared to 23.5%. Additionally, the false negative rates for white defendants were about twice as high than for black defendants, 47.7% compared to 28.0%.

COMPAS assesses recidivism risk using answers to 137 questions that are either directly answered by the defendant or taken from their personal history. These questions target five main areas: criminal involvement, relationships/lifestyles, personality/attitudes, family, and social exclusion [2]. The algorithm outputs a risk score, a number from 1-10, that represents their likelihood of recidivating, with 10 being the most likely to recidivate relative to all others. These risk scores are used by judges to determine anything from bail and early release, to parole and sentencing decisions. In other words, this

algorithm uses demographic information, of which defendants often have little to no control over, to make decisions over the amount of time defendants will spend in cages.

COMPAS sits among the most widely used software of its kind in the country. This algorithm is currently used in criminal sentencing processes in eight states so far, including Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia and Washington [2].

1.1.1 Points of Controversy

The following are some of the issues and questions that have garnered controversy, especially since the release of ProPublica's article.

Constitutional Rights

A big critique surrounding the use of statistics to predict human behavior in the criminal justice system is that it violates constitutional issues. Following, due process of law requires that anytime the government is taking life, liberty, or property from a person, the accused must be able to confront and cross examine any witnesses against them Goldberg v. Kelly [3]. The definition of due process is a bit more conservative in criminal law, but it is there nonetheless. While the computer algorithm, COMPAS, may have very significant influence over a judge's decisions, it has no way of being cross-examined. Northpointe, the for-profit company that created and sells this computer algorithm, does not make it available to the public. Even if it were available, it would not necessarily be accessible to the different players in courtroom. While judges do receive some training on the algorithm, it is hard to say whether its use and limitations are equally clear among judges, and it is also difficult to say how much weight they will actually give a risk score when making sentencing decisions.

Accessibility Issues

What are the ethics behind distributing a simplified result of this algorithm to judges who may not have a strong sense of its limitations and who will use it to make life-altering decisions for others on a daily basis? Northpointe published *A Practitioner's Guide to COMPAS* (2012), which details some of the crimonological theories and framework under which interpretation can

be accomplished to varying degrees. While this document is available to the public, it is unclear whether judges actually receive sufficient training to be able to discern the limits of the algorithm's predictive capacity, especially in light of the particularities of a case.

Role in the Criminal Justice Process

Tim Brennan, co-founder of Northpointe and co-creator of the algorithm, admits that he didn't design COMPAS to be used in sentencing [1]. "I wanted to stay away from the courts," Brennan said, "But as time went on I started realizing that so many decisions are made, you know, in the courts. So I gradually softened on whether this could be used in the courts or not." COMPAS was originally designed to aid in determining treatment and resource allocation to minimize recidivism risk, since Brennan's original intent was on reducing crime instead of punishment. The use of COMPAS has evolved to one of helping determine prison sentences. Similarly, other predictive tools like it are entering every step of the criminal justice process, from bail and early release, to probation and decision making around rehabilitation programs. Some argue that these algorithms should have a restorative goal and should be explicitly used for decision making around probation or rehabilitation services, that they have no place in the sentencing process. Others argue that risk assessment tools help reduce prison and jail populations and should be used in sentencing.

Doomed from the Start?

If the algorithm's training data is biased in and of itself, then COMPAS will treat all subsequent cases in the context of this bias. In light of this, is it fair for an individual's case to predicated on the outcomes of people in similar situations who have gone before? The use of COMPAS is eerily reminiscent of the film *Minority Report* (2002) and the idea of deciding the fate of people based on crimes they have not yet committed.

Compared to What?

Taking ProPublica's analysis on racial biases to be true, a question that's left standing is, even with the inherent racial biases in COMPAS, does it still fare off better than racial biases already present in judges? While COMPAS has a clear and quantified measure of error based on racial bias, what are

the inherent biases and factors that more generally contribute to a *judge's* decision-making process? To what extent are these quantifiable? Danziger et al. [4], for example, showed that judges were the least more likely to make more favorable ruling decisions following their lunch break. While this is an issue that a computer algorithm will not have, what other trade-offs exist between the use or not of COMPAS and other risk assessment tools?

What is the Right Question?

The Bureau of Justice found that 67.8 percent of released prisoners were rearrested within three years of release [Durose et al., 2014]. B. Starr [5] at University of Michigan Law School argues that if the goal is to decrease recidivism rates at a large scale, the question judges should be answering is not "Who is most prone to recidivating?" but rather, "For whom will incarceration minimize recidivism risk the most?" Still, others are focusing not on *if* we can measure human behavior in the criminal justice system, whether it be defendants' recidicism post-release or judges' racial bias in the courtroom, but rather, *should* we even be trying to do this?

1.2 Moving Forward

At the end of the day, these algorithms have already found their way into the prison industrial complex and are presently affecting the flow of human bodies in and out of cages. This paper explores the mathematical workings of algorithms like COMPAS and aims to critically explore questions of the limits of the unquestioned authority typically ascribed to numbers and data that claim to be "objective" and without a political agenda. It seeks to question a system that allows for-profit companies to create, sell, and profit off of algorithms that have potentially life-altering consequences for people. It seeks to conextualize this issue in light of the fact that the history of prisons and slavery in this country are inextricably linked and I would argue, a continuation of each other.

1.2.1 Overview

Ultimately, this paper aims to answer two questions: First, what drives disparate impact across race? More specifically, what affect does biased data

have on this disparate impact? I begin with some data exploration analysis to investigate the way disparate impact behaves with respect to the COMPAS algorithm, outline a simulation that helps to answer the latter question, and end with some concluding thoughts that follow from the simulation results.

Chapter 2

COMPAS Data Exploration Analysis

2.1 Sensitivity Trade-offs

Recent scholarly literature suggests that it is mathematically impossible for statistical risk assessments to have equal rates of both correct and incorrect predictions across a given demographic [6, 7, 8]. This chapter examines this claim further, and investigates the tradeoffs between predictive accuracy and disparate impact using the data of criminal defendants in Broward County, Florida collected by Angwin et al. [1] and used in their analysis.

2.1.1 Disparate Impact

Chouldechova [7] argues that differences in prevalence across groups drive disparate impact, a direct result of the following:

$$FPR = \left(\frac{p}{1-p}\right)\left(\frac{1-PPV}{PPV}\right)(1-FNR)\,, \tag{2.1}$$

where FPR is the False Positive Rate, FNR is the False Negative Rate, PPV is Positive Predictive Value and p(lowercase) is prevalence.

Below, we derive equation 2.1. Let P (uppercase) be the total number of real positive cases in the data, N be the total number of real negative cases, p (lowercase) again be the prevalence rate, FP and TP be the total number of

false positives and true positives, respectively, and FN and TN be the total number of false negatives and true negatives, respectively. Then, we get the following.

Claim: $FPR = \left(\frac{p}{1-p}\right)\left(\frac{1-PPV}{PPV}\right)(1-FNR)$

Proof:

$$FPR = \frac{FP}{N} \tag{2.2}$$

$$= \left(\frac{FP}{N}\right)\left(\frac{P}{P}\right) \tag{2.3}$$

$$= \left(\frac{FP}{N}\right)\left(\frac{P}{P}\right)\left(\frac{TP}{TP}\right) \tag{2.4}$$

$$= \left(\frac{FP}{TP}\right)\left(\frac{TP}{N}\right)\left(\frac{P}{P}\right)\left(\frac{TP+FP}{TP+FP}\right) \tag{2.5}$$

$$= \left(\frac{FP}{TP}\right)\left(\frac{TP}{N}\right)\left(\frac{P}{FN+TP}\right)\left(\frac{TP+FP}{TP+FP}\right)\left(\frac{P+N}{P+N}\right) \tag{2.6}$$

$$= \left(\frac{P}{P+N}\frac{P+N}{N}\right)\left(\frac{FP}{TP+FP}\frac{TP+FP}{TP}\right)\frac{TP}{FN+TP} \tag{2.7}$$

$$= \left(\frac{\frac{P}{P+N}}{\frac{N}{P+N}}\right)\left(\frac{\frac{FP}{TP+FP}}{\frac{TP}{TP+FP}}\right)\left(1-\frac{FN}{FN+TP}\right) \tag{2.8}$$

$$= \left(\frac{\frac{P}{P+N}}{1-\frac{P}{P+N}}\right)\left(\frac{\frac{FP}{TP+FP}}{\frac{TP}{TP+FP}}\right)\left(1-\frac{FN}{FN+TP}\right) \tag{2.9}$$

$$= \left(\frac{p}{1-p}\right)\left(\frac{1-PPV}{PPV}\right)(1-FNR) \tag{2.10}$$

□

In the case of COMPAS results for black and white defendants, Chouldechova makes the following claim, which we prove.

Claim: If $PPV_B = PPV_W$ and $p_B \neq p_W$, then $FPR_B = FPR_W$ and $FNR_B = FNR_W$ cannot both be true.

Proof: Let $PPV_B = PPV_W$ and $p_B \neq p_W$. Then,

$$FPR_B = \left(\frac{p_B}{1-p_B}\right)\left(\frac{1-PPV_B}{PPV_B}\right)(1-FNR_B) \tag{2.11}$$

$$= \left(\frac{p_B}{1-p_B}\right)\left(\frac{1-PPV}{PPV}\right)(1-FNR_B)\,. \tag{2.12}$$

Similarly,

$$FPR_W = \left(\frac{p_W}{1-p_W}\right)\left(\frac{1-PPV_W}{PPV_W}\right)(1-FNR_W) \tag{2.13}$$

$$= \left(\frac{p_W}{1-p_W}\right)\left(\frac{1-PPV}{PPV}\right)(1-FNR_W) \tag{2.14}$$

Case I:Assume $FPR_B = FPR_W$. Then,

$$\left(\frac{p_B}{1-p_B}\right)\left(\frac{1-PPV}{PPV}\right)(1-FNR_B) = \left(\frac{p_W}{1-p_W}\right)\left(\frac{1-PPV}{PPV}\right)(1-FNR_W) \tag{2.15}$$

$$\left(\frac{p_B}{1-p_B}\right)(1-FNR_B) = \left(\frac{p_W}{1-p_W}\right)(1-FNR_W) \tag{2.16}$$

$$\tag{2.17}$$

Since $p_B \neq p_W$, $FNR_B \neq FNR_W$.
Case II:Assume $FNR_B = FNR_W$. Then,

$$1-(FPR_B)\left(\frac{1-p_B}{p_B}\right)\left(\frac{PPV}{1-PPV}\right) = 1-(FPR_W)\left(\frac{1-p_W}{p_W}\right)\left(\frac{PPV}{1-PPV}\right) \tag{2.18}$$

$$1-(FPR_B)\left(\frac{1-p_B}{p_B}\right) = 1-(FPR_W)\left(\frac{1-p_W}{p_W}\right) \tag{2.19}$$

Since $p_B \neq p_W$, $FPR_B \neq FPR_W$.
□

In other words, given the assumption that black and white defendants recidivate at different rates ($p_B \neq p_W$), and given that we want our risk assessment tool to correctly predict recidivism at the same rates across races

($PPV_B = PPV_W$), it is mathematically impossible for our risk assessment tool to have equal rates of incorrect predictions across the two groups (both $FPR_B = FPR_W$ and $FNR_B = FNR_W$ cannot both be true). These results help answer the first motivating question for my thesis: what drives disparate impact? We can see that differences in prevelance rates across any two demographic groups is one driver of disparate impact in their rates of incorrect predictions.

2.2 Tradeoffs in Practice

Following the example of Angwin et al. [1], I subsetted the Broward County dataset to include only defendants whose COMPAS-scored crime was within 30 days of their arrest, who had a COMPAS assessment, whose case was not based on an ordinary traffic offense, and who either recidivated in two years, or had at least two years without recidivating. Finally, because observations were limited for other races, I filtered for black and white defendants only. Using this data, the following analysis was performed.

2.2.1 Statistical Significance of Race

As with Angwin et al. [1]'s analysis, I conducted a logistic regression model with the new subset of data. The model uses sex, age category (less than 25 years old, 25-45, or greater than 45), race (black or white), number of previous offenses, degree of charge (felony or misdemeanor), and whether or not a defendant recidivated within two years following their COMPAS assessment to find the probability of a defendant being labeled high risk by COMPAS. The logistic regression results are given in Table 2.1.

We use this logistic regression to quantify the association between the two categories of race, black and white, and the binary response variable for high COMPAS risk score. Calculating the odds ratio (OR) would suffice, and would in fact be relatively simple to compute. However, since results in terms of probabilities are usually more intuitive to understand, we use relative risk (RR) instead. According to Zhang and Yu [9], relative risk can be calculated as follows:

$$RR = \frac{OR}{(1 - P_0) + (P_0 * OR)}, \tag{2.20}$$

where P_0, in this case, indicates the probability of receiving a high COMPAS risk score for white defendants, P_1 indicates the probability of receiving a high COMPAS risk score for black defendants, and

$$OR = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}.$$

For small levels of incidence ($< 10\%$) across different groups, OR is a good approximation for RR [9]. This is the case because as P_0 and P_1 approach 0, OR approaches $\frac{P_1}{P_0}$, which is equivalent to RR. However, as the values of P_0 and P_1 increase, OR deviates more and more from $\frac{P_1}{P_0}$ and OR becomes an increasingly worse approximation for RR. Since P_0 and P_1 are relatively large for black and white defendants, we use Zhang and Yu [9]'s derivation of RR. Note that RR also accounts for other covariates while OR does not.

The coefficient for being African American, β_1, was 0.47992 and was statistically significant at the $\alpha = .0001$ level. Similar to Angwin et al. [1], I first calculated the logistic of the intercept, where the reference category for race represented by the intercept is white defendants. I then calculated

$$RR = \frac{P_1}{P_0} = \frac{e^{\beta_4}}{1 - logistic(intercept) + (logistic(intercept)e^{\beta_4})}, \tag{2.21}$$

and obtained a value of 1.456707. The calculation for RR in 2.21 follows from 2.20 and is equivalent to 2.20. Consequently, the result of 1.456707 indicates that, compared to white defendants, black defendants are about 46% more likely to receive a high score controlling for sex, age, the seriousness of their crime, previous arrests, and future rearrest.

How would race-based thresholds affect the observed relative risk for being labeled high risk across races? To answer this question, I created a new response variable for black defendants of decile risk scores lowered by 1. A new logistic regression model with the aforementioned updated response variable and the same predictor variables as my previous model is given in Table 2.2.

	Coef.	SE
(Intercept)	-1.53^{***}	(0.08)
gender_factorFemale	0.32^{***}	(0.08)
age_factorGreater than 45	-1.37^{***}	(0.11)
age_factorLess than 25	1.24^{***}	(0.08)
race_factorAfrican-American	0.48^{***}	(0.07)
priors_count	0.27^{***}	(0.01)
crime_factorM	-0.33^{***}	(0.07)
two_year_recid	0.71^{***}	(0.07)
Num. obs.	5114	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2.1: Logistic Regression Model for Unaltered COMPAS Classification

The coefficient for black defendants, β_4, in this case is -0.09340 and is not statistically significant. The intercept for my new logistic regression model, β_0, is -1.50837. The intercept represents the coefficient for white defendants, given that all other coefficients are 0. In other words, the reference category represented by the intercept is white, male, defendants between the ages of 25 and 45 who have 0 priors, committed felonies, and did not recidivate within two years. I found the logistic of the intercept by calculating $\frac{e^{\beta_0}}{1+e^{\beta_0}}$. Similar to the equation in 2.21, I calculated the following,

$$RR = \frac{P_1}{P_0} = \frac{e^{\beta_4}}{1 - logistic(\beta_0) + (logistic(\beta_0)e^{\beta_4})},$$

and obtained a value of 0.9257861. This indicates that if every COMPAS score for Black defendants is subtracted by 1, White defendants are about 7% more likely than Black defendants to receive a high score, controlling for the seriousness of their crime, previous arrests, and future criminal behavior. Moreover, race with a coefficient of -0.09 and a p-value of 0.19, is no longer a significant predictor for COMPAS score. This means that raced based thresholds for determining low-risk from high-risk defendants can in fact remove the predictive ability of race as determinant of a defendant's COMPAS score. In other words, if we want both black and white defendants to be equally as likely to receive a high score for the COMPAS assessment, then we need to shift all the COMPAS risk scores for black defendants by one in the negative direction or by all the COMPAS risk scores for white defendants by one in the positive direction. This would consequently change the threshold that COMPAS uses to differentiate low-risk defendants from

high-risk defendants. For example, if a risk score of 5 previously deemed black and white defendants as high risk, then adjusting the threshold would mean that black defendants now need a risk score of 6 to be deemed high risk while the threshold for whites stays the same. Equivalently, if the threshold of 5 previously deemed both groups as high risk, white defendants now only need a risk score of 4 to be deemed high risk, while the threshold for blacks stays the same. If we were to set up race-based thresholds for COMPAS classification in this manner, race would no longer be significant in predicting COMPAS score.

	Coef.	SE
(Intercept)	-1.53^{***}	(0.08)
gender_factorFemale	0.32^{***}	(0.09)
age_factorGreater than 45	-1.37^{***}	(0.11)
age_factorLess than 25	1.26^{***}	(0.08)
race_factorAfrican-American	-0.09	(0.07)
priors_count	0.26^{***}	(0.01)
crime_factorM	-0.36^{***}	(0.07)
two_year_recid	0.73^{***}	(0.07)
Num. obs.	5114	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2.2: Logistic Regression Model for COMPAS Risk Scores with 1 Subtracted from Every Black Defendant's Risk Score

As an alternative modeling strategy, I also created an additional logistic regression model that uses the sex, age category, race, number of previous offenses, degree of charge, and COMPAS classification (low or high risk) to find the probability of a given defendant actually recidivating within two years. Note that the risk score classification threshold deems defendants with risk scores above 4 as high risk of recidivating. The model is presented in Table 2.3. The results of this model indicate that race is not a significant factor in determining whether or not an individual recidivates, given the other variables in the model. However, the defendant's risk category classification based on COMPAS risk scores is significant. Based on Table 2.1 and Table 2.3, notice that while race is a significant predictor of COMPAS classification and COMPAS classification is a significant predictor of actual recidivism within two years, race is not a significant predictor of actual re-

cidivism within two years, given COMPAS score (and the other variables). This is an interesting finding that needs further investigation.

	Coef.	SE
(Intercept)	-0.77^{***}	(0.07)
gender_factorFemale	-0.43^{***}	(0.08)
age_factorGreater than 45	-0.55^{***}	(0.11)
age_factorLess than 25	0.55^{***}	(0.08)
race_factorAfrican-American	0.05	(0.07)
priors_count	0.14^{***}	(0.01)
crime_factorM	-0.13^{*}	(0.07)
score_factorHigh	0.74^{***}	(0.07)
Num. obs.	5114	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2.3: Logistic Regression Model for Actual Recidivism

2.2.2 Positive Predictive Value v. Disparate Impact

What affect does changing COMPAS classification thresholds for each race have on the false positive rates, false negative rates, and positive predictive values for COMPAS predictions? Is there functional relationship as COMPAS classification thresholds change? To answer these question, I tabulated COMPAS predictions (low risk or high risk) with defendant's two-years recidivism. In other words, I computed confusion matrices between COMPAS predictions and actual recidivism, and did this separately for black and white defendants, for separate different race-based thresholds. According to Bloomberg et al. [10], the current classification thresholds for the COMPAS risk assessment are as follows: a risk score of 4 or below classifies defendants as low risk, and a risk score of 5 to 7 as medium risk, and a risk score higher than 7 as high risk. Similar to Angwin et al. [1], I set all risk scores above 4 to be high risk, and any risk scores below 4 to be low risk. I do so to simplify the analysis by analyzing a binary response as opposed to a three-category predictor.

Consider Table 2.4. To obtain the results from this table, I computed the following. Using the data at hand, I tabulated COMPAS predictions and actual two-year recidivism results for black defendants. The results for false positive rate, false negative rate, and positive predictive value for this confusion

matrix are indicated in the column titled "Black" of Table 2.4. Additionally, I created a new variable for black defendants that took their risk score and subtracted it by 1. Under these new risk scores for black defendants, I then created a new dummy variable that classified their new risk scores under the same thresholds, low risk if the risk score is 4 or less and high risk if the risk score is 5 or above. Thus, although the COMPAS classification threshold technically remained the same, the risk scores for black defendants were treated less severely than whites, by 1 score number. I then tabulated the actual new COMPAS predictions given the change in risk scores for black defendants and the same actual two-year recidivism values as before, to get a new confusion matrix. The results for this are found in Column 2, titled "Black(-1)," of Table 2.4. I created new variables for black defendants that took their risk score and subtracted it by 2, 3, and 4, respectively, and for each, repeated the process of tabulating COMPAS predictions under these new risk scores against two-year recidivism. These results are found in the last three columns of 2.4.

Consider Table 2.5. I repeated the same process for white defendants as I did for black defendants. Instead of incrementally subtracting their risk scores by 1, however, I incrementally added 1 to their risk scores, and computed their corresponding confusion matrices. The results are as follows.

	Black	**Black(-1)**	**Black (-2)**	**Black (-3)**	**Black (-4)**
FPR	41.44	30.60	22.18	13.55	8.13
FNR	28.38	37.99	49.25	61.83	74.77
PPV	67.16	70.60	73.05	76.85	78.61
p	54.23	54.23	54.23	54.23	54.23

Table 2.4: Confusion Matrix Results for Black Defendants

Given the original COMPAS risk scores, the confusion matrix results for black and white defendants show that the positive predictive value is more disparate than both Northpointe and ProPublica indicated, with 60.88% for white defendants and 67.16% for black defendants, as opposed to the approximate 61% mentioned by Angwin et al. [1]. This is due to the fact that I filtered the dataset in a slightly different fashion. There were 164 defendants who recidivated *after* the two-year period following their COMPAS assess-

	White	White(+1)	White(+2)	White(+3)	White(+4)
FPR	21.64	32.71	44.91	62.08	100
FNR	49.64	37.71	27.74	15.57	0
PPV	60.88	56.02	51.83	47.63	40.08
p	40.08	40.08	40.08	40.08	40.08

Table 2.5: Confusion Matrix Results for White Defendants

ment. While Angwin et al. [1] did not filter these cases out, I did before beginning my analysis, which explains the slightly different results.

Consider Column 3 from Table 2.4 and Column 1 from Table 2.5. Notice that the false positive rates for black and white defendants are approximately equal, 22.18 and 21.64, respectively. Similarly, the false negatives are approximately equal, 49.25 for black defendants, and 49.64 for white defendants. By taking every risk score received by a black defendant, subtracting it by two, and classifying under the same thresholds for both black and white defendants (≤ 4 as low risk and > 4 as high risk), the FPR and FNR balanced out across the groups. However, there is now a larger disparity in PPV. The PPV for black defendants is 73.05 while the PPV for white defendants is 60.88. As we worked toward balancing out FPR and FNR across the races, the disparity for PPV increased. This is an example of the equation derived by [7] at play. Given the different prevalence rates, 54.23% for black defendants and 40.08% for white defendants, it is impossible to balance FPR and FNR rates without creating a further disparity in PPV.

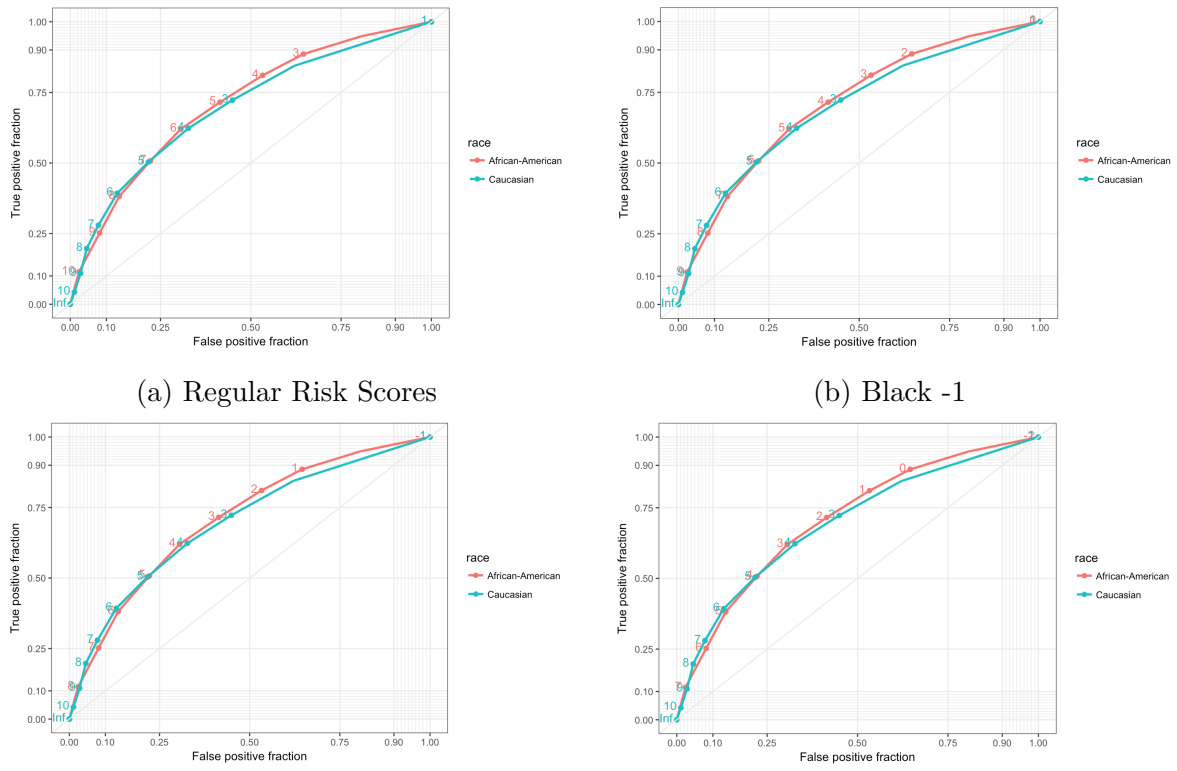
2.2.3 ROC Curves Analysis

ROC Curves are graphical representations of the specificity-sensitivity tradeoffs between different threshold values, in this case, for a binary classifier. The x-axis for these plots represents the false positive rate, or the probability of false alarm, given by (1−specificity). Specificity is calculated as total number of true negatives over total number of negatives($\frac{TN}{N}$). The y-axis represents the true positive rate, or sensitivity, which is given by total number of true positives over total number of positives ($\frac{TP}{P}$). The line $y = x$ represents the specificity-sensitivity values for a random classifier, where specificity and

sensitivity are always equivalent. Anything above the line $y = x$ means this classification threshold is better than a random classifier and anything below this line means it is worse. Additionally, the best classifier will have 100% sensitivity, and 100% specificity. In other words, it will always detect when needed, and it will never be a false alarm. This is the point (0,1) on the ROC graph. Thus, the point closest to (0,1) will be the best threshold for a given classifier.

I created an ROC curve for each race, using COMPAS risk scores and two-year recidivism results. This ROC curve is given in plot (a) of Figure 2.1. Consider the point labeled "4" on the ROC curve for white defendants, and the point labeled "6" on the ROC curve for Black defendants. Notice that they are the closest points to each other. This is saying that a threshold of 6 for black defendants will have about the same probability of detention and probability of false alarm as a threshold of 4 for white defendants. Notice also that these two points are some of the closest to the point (0,1). A threshold of 4 for white defendants and a threshold of 6 for black defendants, then, are the best thresholds if we want to maximize probability of detection and minimize probability of false alarm for each race.

Say that we wanted to balance these thresholds. We want for a threshold of 4 for white defendants to have the same TPR and FNR as the threshold of 4 for black defendants. I made a new variable that took every risk score for black defendants and subtracted it by 1. I then used this variable with new risk scores along with the binary variable for two-year recidivism and plotted a new ROC curve for each race, given by plot (b) in Figure 2.1. I repeated this same process, but instead subtracted each risk score for black defendants by 2 and 3, respectively, and plotted the ROC curves given these new risk scores for black defendants next to the ROC curves for white defendants, whose risk scores were fixed. These plots are given in parts (b) and (c) of Figure 2.1, respectively. Notice from plot (b) that subtracting black defendants' risk scores by 1, we get that a threshold of 5 for black defendants is approximately equal to a threshold of 4 for white defendants. Moreover, from plot (c), we are finally able to balance out the thresholds by race. Plot (c) shows that subtracting every risk score for black defendants by 2 results in approximately equal thresholds across races. Thus, in order for the probability of detection and the probability of false alarm to be approximately



(a) Regular Risk Scores

(b) Black -1

(c) Black -2

(d) Black -3

Figure 2.1: ROC Curves by Race

equal for COMPAS classification, the thresholds need to be different across races. If a white defendant is considered high risk with a risk score of 5 or above, a black defendant must be considered high risk with a risk score of 7 or above, in order for COMPAS classification to have the same meaning for both races. Consider the points labeled "4" in the ROC curves for both black and white defendants in plot (a). Notice that the "4" on the ROC curve for black defendants is much higher than that of white defendants, both interms of FPR and TPR. Thus, continuing to use the current threshold of 4 for both groups means that the COMPAS assessment will predict future recidivism correctly for a higher proportion of black defendants than whites at the expense of mislabeling more black defendants as high risk. In order to remedy this disparate impact, either the classification threshold must go up by two for black defendants, or down by two for white defendants.

Chapter 3

Simulating the Effects of Over-policing Practices on COMPAS Assessments

3.1 Purpose and Outline of Simulation Study

Through the following simulation exercise, we want to get closer to understanding whether COMPAS is doomed from the start. Does bias in data drive disparate impact? To answer this question, we use this simulation exercise to test whether oversampling a disadvantaged group leads to a feedback loop that drives up the group's false positive rates. We can think of oversampling in this simulation study as the equivalent of implementing policies that target and over-police disadvantaged groups. Thus, we ask whether a cycle of oversampling from a disadvantaged group and subsequently recalibrating a predictive model on new oversampled data perpetuates and intensifies disparate impact among groups.

I define a *disadvantaged group* to be a group that, compared to their counterparts, will be mislabeled as high risk at a higher rate by the COMPAS Assessment. I base these groups based on Angwin et al. [1] results. Angwin et al. [1] found that there was a significant difference in COMPAS scores across race, age, and sex, namely that black, young ($<$ 25 years old),and female defendants were more likely to receive a higher COMPAS assessment than white, middle aged ($25 - 45$ years old), and male defendants, respec-

tively. To obtain these results, they created a logistic regression model that used sex (male or female), age (young, < 25 years old, middle aged, $25 - 45$ years old, or old, > 45 years old), race (White, African-American, Asian, Hispanic, or other), number of prior arrests (a continuous variable), type of crime committed (misdemeanor or felony), and whether or not a defendant recidivated within two years following the date of their COMPAS assessment, to predict COMPAS classification (low risk or high risk of recidivating). Using the coefficients from this model, they calculated the logit of the intercept, which represents, white, middle aged, male defendants and compared it to the logit of the coefficients for black, young, and female defendants, respectively. They found that black defendants were 45% more likely than white defendants to receive a higher score, accounting for sex, age, number of prior arrests, type of crime committed, and two year recidivism. Similarly, they found that young defendants were about 2.5 times as likely to get a higher score than middle aged offenders, holding all other variables constant. Moreover, holding all other variables constant, female defendants were 19.4 percent more likely to get a higher score than men. Thus, for the data we are working with, I consider the disadvantaged groups within race, age, and, gender variables to be black, young (less that 25 years old), and female, respectively[1].

For this simulation, assume that we draw from the same population every time. The skeleton for our simulation process is as follows:

1. Round 1
    (a) Simulate sample data with little to no bias from a reasonable population
    (b) Create a logistic model that predicts recidivism within two years using this sample data
    (c) Simulate new sample data and predict using the model created in (b)
    (d) Find false positive and false negative rates of predictions for this round
2. Round 2
    (a) Simulate sample data with plenty of bias from the same reasonable population

(b) Create a logistic model that predicts recidivism within two years using this sample data with oversampling

(c) Simulate new sample data (still with oversampling) and predict using the model created in (b)

(d) Find false positive and false negative rates of predictions for this round

More specifically, the simulation exercise is as follows:

1. Using data obtained from [1], call it $Data_0$, create a logistic regression model that predicts actual recidivism within two years of the COMPAS assessment, based on type of crime arrested for (misdemeanor or felony), age category (less than 25 years old, 25-45 years old, or greater than 25 years old), race (black or white), gender (male or female), COMPAS assessment (low risk or high risk), and number of priors. Call this logistic regression model $Model_0$. Use $Model_0$ to stand-in for your reasonable population. In other words, $Model_0$ stores the relationships found in $Data_0$ so that every time we draw a new sample from this model, we draw from a population with reasonable variable values and relationships.

2. Round 1

   (a) Simulate a sample from the population using $Model_0$, and set the proportion of black defendants to be equal to 0.6, about the same as Angwin et al. [1]'s data sample. Call this set of simulated X_i's, $predictors_{1b}$. Use $Model_0$ to make predictions for $predictors_{1b}$. Use the probabilities of recidivating given by these predictions to draw some truth for $predictors_{1b}$ from a binomial distribution, according to those probabilities. Call it $Truth_{1b}$ and say that $Truth_{1b}$ represents whether or not each defendant from the simulated sample ended up recidivating within two years following their risk assessment. I use a subscript of "b" for any simulated data related used in building a new model. Moreover, I use a subscript of "p" for any simulated data on which this new model will predict.

   (b) Create a new logistic regression model that predicts two-year recidivism using type of crime arrested for (misdemeanor or felony),

age category (less than 25 years old, 25-45 years old, or greater than 25 years old), race (black or white), gender (male or female), and number of priors. Call it $Model_1$.

(c) Draw a new sample from the population using $Model_0$ and call it $predictors_{1p}$. Create some truth for this simulated sample of defendants similarly to the truth created in (a). Call it $Truth_{1p}$

(d) Make predictions on $predictors_{1p}$ using $Model_1$. Tabulate these predictions against $Truth_{1p}$ to calculate the false negative and false positive rates for black and white defendants.

3. Round 2

(a) Simulate a sample from the population using $Model_0$ and oversample for black defendants, where black defendants make up 90% of the sample. Call this sample $predictors_{2b}$. Simulate some truth for this sample in the same way it was simulated in steps (1a) and (1c). Call it $Truth_{2b}$.

(b) Create a new logistic regression model that predicts two-year recidivism using the same predictors as $Model_1$. Call it $Model_2$.

(c) Draw a new sample from the population and call it $predictors_{2p}$. Generate some truth, $Truth_{2p}$ for this new sample.

(d) Make predictions on $predictors_{2p}$ using $Model_2$. Tabulate these predictions against $Truth_{2p}$ to calculate the false negative and false positive rates for black and white defendants.

4. Compare results for Rounds 1 and 2.

Through this simulation exercise, two models were built. $Model_1$ was built from data with little to no oversampling, or with little to no bias. $Model_2$, however, was built using biased data, with 90% of the data being black defendants. We now ask, how well does $Model_2$ do, compared to $Model_1$?

3.2 Results

3.2.1 Oversampling Drives Disparate Impact!

Notice the last column on 3.1 and 3.2. For both rounds of the simulation, the FPR for black defendants was higher than their FNR. On the other hand, the

FPR for white defendants was lower than their FNR. For the $FPR - FNR$ column in 3.1 and 3.2 then, we know that this difference measures how much higher FPR is than FNR for black defendants, and how much higher FNR is than FPR for white defendants. Consequently, this difference is a measure of disparate impact. In the case of the simulation described above, both models had disparate impact, Model 2's disparate impact was worse. With Model 1, the FPR and FNR for black defendants were fairly similar, with a difference of about 0.05. With Model 2, however, the difference between FPR and FNR is now higher, at about 0.09. This means that for Model 2, black defendants are mislabeled high risk more often, relative to how often they are mislabeled low risk. On the other hand, the FPR for white defendants was lower than their FNR for Model 1 by -0.2, and for Model 2 by -0.27.

From these results, we can see that oversampling does in fact worsen disparate impact, since for both races, the gap between FPR and FNR increases from Model 1 to Model 2.

Notice from Tables 3.1 and 3.2 however, that predictive accuracy does seem to improve. 57% of black defendants were given correct predictions on future recidivism for Model 1 while 72% were correctly predicted for Model 2. For white defendants, about 56% were correctly predicted by Model 1 while about 66% were correctly predicted for Model 2. While predictive accuracy improved, disparate impact worsened. Whether this inverse relationship is always the case is a claim to be examined for future research.

	FPR	**FNR**	**PPV**	**Accuracy**	**TP+TN**	**Total**	**FPR-FNR**
Black	0.46	0.41	0.61	57%	1730	3029	0.05
White	0.36	0.56	0.45	56%	1104	1971	-0.20

Table 3.1: Model 1 Results (60% Black Defendants)

	FPR	**FNR**	**PPV**	**Accuracy**	**TP+TN**	**Total**	**FPR-FNR**
Black	0.32	0.24	0.74	72%	3237	4469	0.09
White	0.23	0.50	0.62	66%	351	531	-0.27

Table 3.2: Model 2 Results (90% Black Defendants)

Chapter 4

Conclusion

These results indicate that biased data, where bias is defined as oversampling of a group, may in fact impact the false positive rates and false negative rates of a model. In the context of the COMPAS risk assessment and risk assessment tools in the criminal justice system more generally, these results present beg some pressing questions. To what extent is biased data influencing the disparate impact observed in COMPAS and presumably, other risk assessment tools currently being used in the criminal justice system? Will obtaining a less biased training set for these algorithms make them more equitable? Is this even possible? If so, what are some ways to obtain a less biased training set?

As previously mentioned, we can take "oversampling" in this simulation to mean implementing policies that target and overpolice a disadvantaged group, in this case, black people. Thus, we can see that policy changes in the way we collect data will impact the way these algorithms perform, both in the way they predict right and in the way they make mistakes.

Algorithms will learn, relearn and spit back complex patterns that they see in the data you give them. So as long as the data is biased, the algorithm will be too. These results affirm the pressing need for data scientists to be more critical of the data used in training algorithms, especially in the context of the criminal justice system, where algorithms have a drastic impact on people's lives. To do so, we need to be critical of the systems lead us to this biased data.

Angela Davis said, "The prison... functions ideologically as an abstract site into which undesirables are deposited, relieving us of the responsibility of thinking about the real issues afflicting those communities from which prisoners are drawn in such disproportionate numbers... It relieves us of the responsibility of seriously engaging with the problems of our society, especially those produced by racism and, increasingly, global capitalism (16)."

I believe algorithms play a similar role of distracting us from the real problem, which is the criminal justice system as a whole and the histories that precede it. More and more, algorithms are entering every step of the criminal justice system, so the work of understanding and analyzing them is especially important. However, we cannot get to the heart of the issue, the much more fundamental problem of the United States' historical amnesia and moral bankruptcy with respect to race and other forms of human classification, by working insularly. To do the work of analyzing the systems that create this biased data, and improve these algorithms, if that's what we want to do, we need more people from different disciplines and backgrounds at the table. We need more than just data scientists and criminologists influencing the creation and use of these algorithms. We need historians, and we need sociologists, and we need formerly incarcerated individuals to all be in conversation to be able to address these issues more comprehensively.

4.1 Further Research

One of the troubling results from my data exploration was that race was significant when two-year recidivism predicted COMPAS classification, but race was not significant when COMPAS classification predicted two-year recidivism. This is one clear point for further investigation. For future reaseach, it would be interesting to examine the intersection of racism and ageism in risk assessments.

The research process could also be fine-tuned to have a more direct comparison between the two models. Notice from Tables 3.1 and 3.2 that the accuracy increases for both races from Model 1 to Model 2. In the future, it would be helpful to choose a different threshold than 0.5 to get a model that is worse at predicting, with approximately the same accuracy as Model 1. This way, the other results in Tables 3.1 and 3.2 can be compared having

balanced for accuracy across the two models.

One of the questions that this project left me continuing to ask is where is there room for interdisciplinary collaboration for the process of creating and implementing risk assessment tools. For example, could oral histories shed new light to why a given county has an overrepresentation of one group of people, and thus lead to strategies for reducing bias in the data, or in other words, alleviating this overrepresentation? Are there policies in place that disproportionately target communities of color? What do sociologists and public policy analysts have to say about the ways these policies are influencing the rates of incarceration and the ways to account for them in the risk assessment tool?

Most importantly, what do formerly incarcerated individuals have to say about recidivism? Risk assessment tools in criminal sentencing function may serve a punitive purpose and assume that a higher risk of recidivism necessitates longer incarceration. How does longer incarceration affect How their life trajectories? If preventing recidivism is the goal, what are the needs of formerly incarcerated individuals?

These are some examples of ways collaboration between people of different disciplines and backgrounds could transform the ways we approach the development, implementation, and understanding of risk assessment tools in the criminal justice system.

Bibliography

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 20(S2), 2016.

[2] Electronic Privacy Information Center. Algorithms in the criminal justice system, 2017. URL `https://epic.org/algorithmic-transparency/crim-justice/`.

[3] Goldberg v. Kelly. 397 U.S. 254. 1970.

[4] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 108:6889–92, 04 2011.

[5] Sonja B. Starr. Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, 66:803–872, 04 2014.

[6] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016. URL `http://arxiv.org/abs/1609.05807`.

[7] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv e-prints*, October 2016.

[8] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016. URL `http://arxiv.org/abs/1610.02413`.

[9] Jun Zhang and Kai F. Yu. What's the relative risk? a method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*, 280(19), 11 1998.

[10] Thomas Bloomberg, William Bales, Karen Mann, Ryan Meldrum, and Joe Nedelec. Validation of the compas risk assessment classification instrument. 2010.