Machine Learning Bias

Pages: 31

Table of contents

1 Introduction

The inevitable robot takeover: we've seen it in novels, TV, conspiracy theories, and more. The image of artificial intelligence (AI) gaining sentience and overthrowing their human creators is one that is familiar to many. We can rest easy, knowing that AI is not as intelligent as sometimes thought, as shown by the neural network created by Janelle Shane to create new ice cream flavors. While eating a "Roasted Beet Pecans" or "Carrot Beer" ice cream cone would certainly be a unique experience, should we perhaps be less concerned with AI taking over the world, or at least the job market, and more concerned with AI making mistakes? Instead of a hyper-intelligent robot takeover, should we be more concerned about AI's lack of intelligence?

In Figure 1.1 we see Google Translate's machine learning translation hard at work. This model has learned from thousands of translations that real people have made, which gives it the ability to translate the ever-changing slang and idioms that real humans use. However, it also learns from the biases that real humans have. When given the sentence "Min kæreste er chef," essentially meaning "My significant other is the boss," the AI behind Google Translate uses

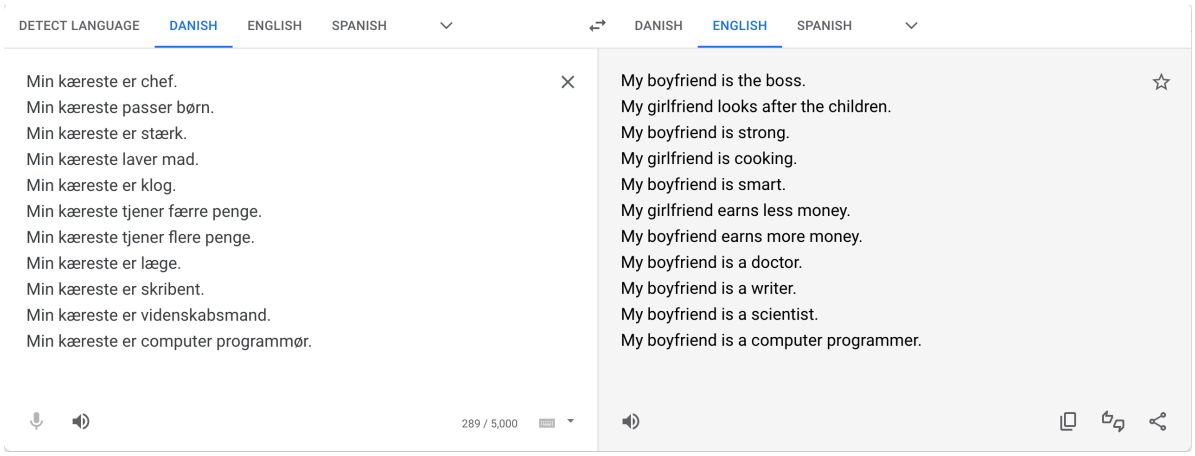

Figure 1.1: Danish to English

its knowledge of the relationship between Danish and English to come up with "My boyfriend is the boss." The translations shown Figure 1.1 are one example of machine learning bias, which according to Mary K. Pratt, "is a phenomenon that occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process." According to IBM, "machine learning is a branch of AI [...] which focuses on the use of data and algorithms to imitate the way that humans learn [...]." Machine learning algorithms create a model based on training data in order to generate predictions without being explicitly programmed how to do so. These models can be biased by gender, race, age, or more. I am a woman, so this bias could directly affect me. The last 4 translations (doctor, writer, scientist, and computer programmer) are all jobs that I have wanted to do at some point in my life, and the machine learning model behind Google Translate has decided are jobs for men. What if instead of a translation algorithm, this was an algorithm that was meant to analyze interview videos and decide who should be hired? Machine learning bias might lead to me not being hired for my dream job, or affect me in countless other ways.

To gain a deeper understanding of machine learning bias, I will be analyzing NorthPointe's COMPAS recidivism algorithm. COMPAS stands for Correctional Offender Management Profiling for Alternative Sanction and it is an algorithm for calculating the risk of an offender recidivating in the next two years. According to the National Institute of Justice, "Recidivism is measured by criminal acts that resulted in rearrest, reconviction or return to prison with or without a new sentence during a three-year period following the person's release." Recidivism is a measure of whether or not a person has reentered the justice system; in an ideal world, it is a measure of whether or not someone has recommited. COMPAS uses many data points (parents' incarceration, previous criminal record, etc.) to calculate risk score, but does not use explicitly use race. I will be analyzing the COMPAS algorithm to see if it is biased by race. If it is biased by race, it could mean that thousands of BIPOC defendants around the country are subjected to longer wait times, higher bail, and harsher sentences.

I will be using a data set collected by ProPublica using records from Broward County, Florida. Florida has strong open-records laws, so ProPublica was able to get both COMPAS scores and criminal records for thousands of people. Some people were later removed, as discussed in the "How We Acquired The Data" section of this article. That left me with 7,214 cases that include a race, decile risk score (1 being low risk, 10 being high risk), a text risk score (1-4 is "Low," 5-7 is "Medium," and 8-10 is "High"), and the truth of whether they recidivated, i.e. had "a criminal offense that resulted in a jail booking and took place after the crime for which the person was COMPAS scored," within two years or not. The data set includes a race variable, which includes African-American, Asian, Caucasian, Hispanic, Native American, and other.

I have grouped these into Black and Indigenous People of Color (BIPOC), which includes anyone with their race identified as African-American or Native American, other people of color (Other POC), which includes anyone identified as Asian, Hispanic, or other, and White, which includes anyone with their race identified as Caucasian. In the United States, all people of color have a history of being oppressed by systemic racism, and it is widely agreed that Black and Indigenous people have been those most disproportionately harmed by the American police and justice system. I have decided to create these three groupings for succinctness and to reflect the varying levels of harm these systems cause.

2 Data

Table 2.1: Percentage recidivism

Race	Risk	Percentage that Recidivated
BIPOC	High	72%
BIPOC	Low	35%
BIPOC	Medium	55%
Other POC	High	62%
Other POC	Low	29%
Other POC	Medium	51%
White	High	71%
White	Low	29%
White	Medium	54%

Table 2.1 shows how often people recidivated based on their text risk score. We can see that a larger percentage of BIPOC individuals recidivated across all risk categories.

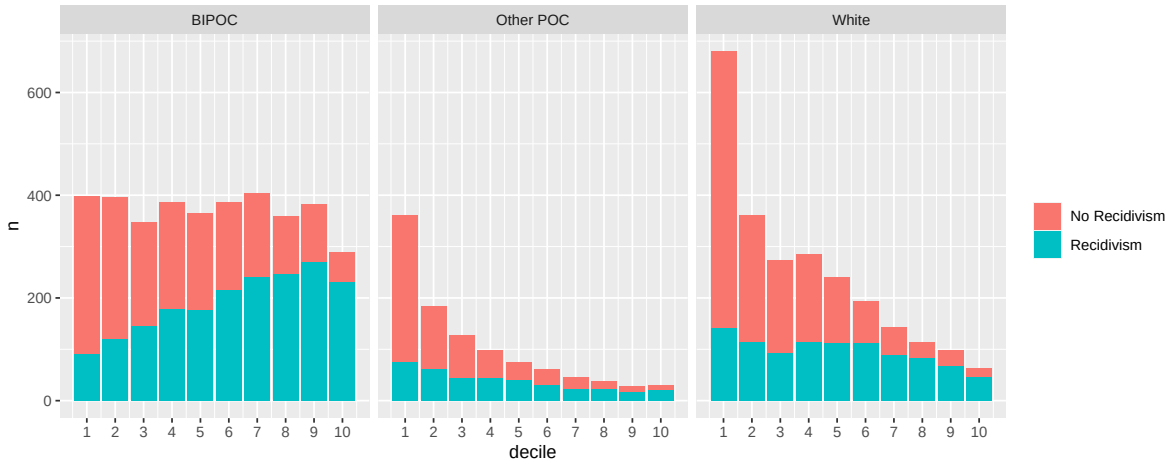

Figure 2.1: Bar chart showing distribution of COMPAS risk score deciles

Figure 2.1 shows us that while the distributions of decile risk scores of both white and other POC individuals are positively skewed, the scores of BIPOC individuals are more evenly distributed. We can also see that the number of people who actually recidivated in the other POC and white groups decreases as the recidivism risk increases, but for the BIPOC group it increases. Does this mean that COMPAS is more accurate for BIPOC people? We can see in Figure 2.2 that the percentage of people who recidivated for any given risk score is fairly similar for all races. COMPAS marks more BIPOC individuals as high risk, but this does not seem to make it significantly more or less accurate for any given racial group.

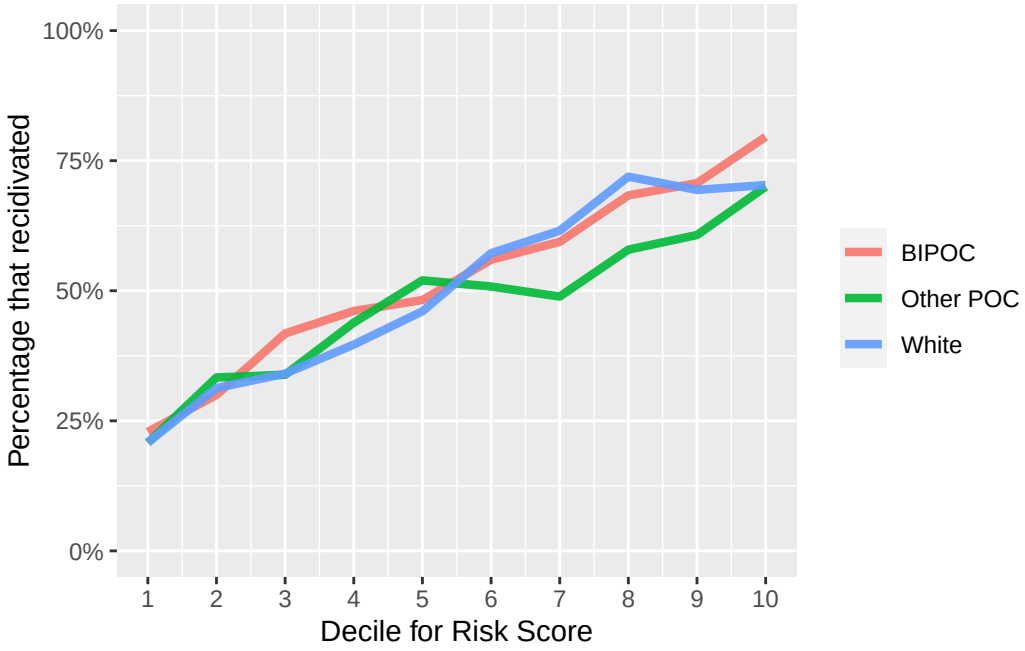

Figure 2.2: Line chart showing percentage of recidivism v. risk score by race

Table 2.2: Base rate recidivism by race

Race	Percentage that Recidivated
BIPOC	51.5%
Other POC	35.8%
White	39.4%

Table 2.3: Number of arrested individuals by race

Race	Number of Arrested Individuals
BIPOC	3714

Race	Number of Arrested Individuals
Other POC	1046
White	2454

Why does COMPAS mark more BIPOC individuals as high risk? Table 2.2 and Table 2.3 show us that more BIPOC individuals are initially arrested, and a larger percentage of arrested individuals go on to recidivate. More BIPOC enter the justice system, and BIPOC individuals reenter the justice system at a higher rate. Therefore, the COMPAS algorithm has learned that marking more BIPOC individuals as high risk would seem to be more accurate, but does that create machine learning bias?

3 Confusion Matrix Calculations

One way to analyze the bias of this algorithm is to create a confusion matrix. A confusion matrix is a 2x2 matrix that compares the prediction of a test or algorithm with the truth. In this case, I will be comparing the COMPAS algorithm's prediction of an individual's recidivism within two years with whether they were actually rearrested within two years. I will be using the COMPAS's risk scores of high, medium, and low. Most individuals are marked as low risk, and so for simplicity and to make the categories more even, I will combine high and medium risk into a "predicted to recidivate" category, while any individual marked low risk is predicted to not recidivate.

As we can see in Figure 3.1, there are four possible categories an individual can land in. They can be predicted to recidivate, and then recidivate (2035 people), which is called a true positive (TP). If they are predicted to recidivate, and then do not recidivate (1282 people), it is called a false positive (FP). If they are predicted to not recidivate, and do not recidivate (2681 people), it is a true negative (TN), and if they are predicted to not recidivate, and then do recidivate (1216 people), it is called a false negative (FN).

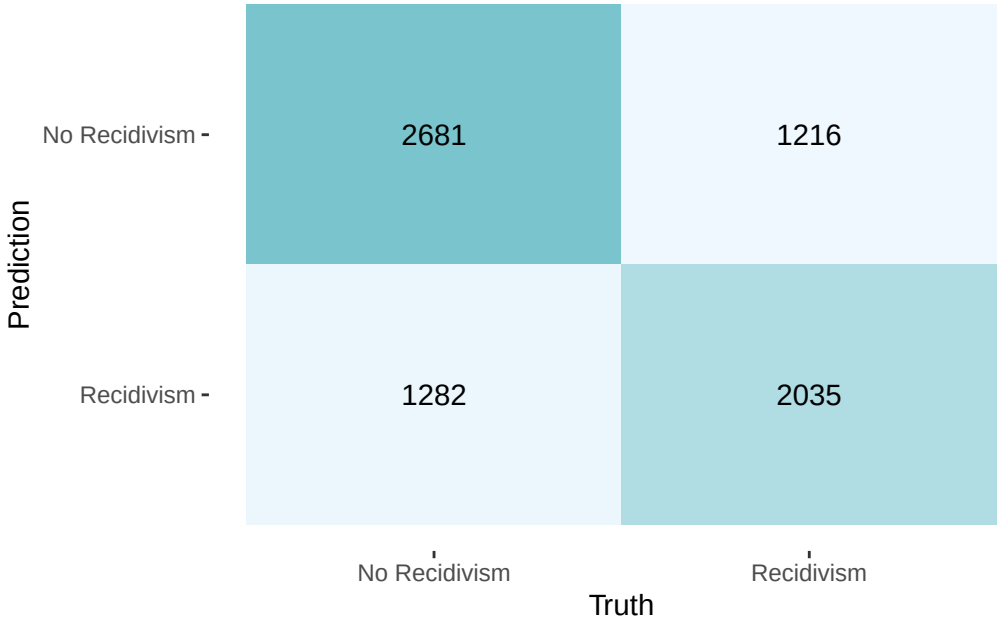

Figure 3.1: Confusion matrix for all individuals

In our confusion matrix, the darkest shaded regions have the most individuals in them. A perfect algorithm or test would place all individuals into the TN and TP category, in this case the top left and bottom right squares. We can see in Figure 3.1 that the two darkest regions are TN and TP, which means that the COMPAS algorithm is right more often than not. However, over 2,400 individuals were still predicted incorrectly.

We can see in Figure 3.2 that the confusion matrices for different races look significantly different. Most notably, in all other confusion matrices, the most common classification is TN, while in the BIPOC confusion matrix, the most common one is TP. You can also see that BIPOC individuals were more likely to be predicted to recidivate than not, while all other individuals are more likely to be predicted to not recidivate. Finally, we can see that in the

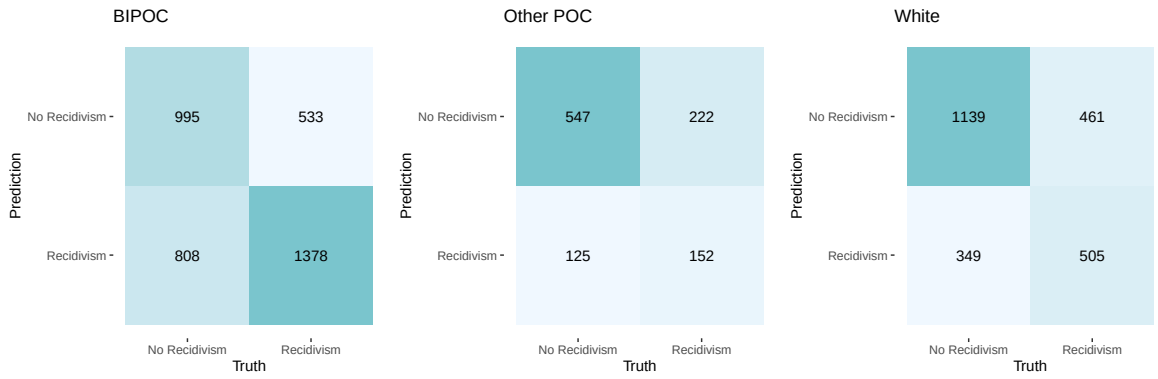

Figure 3.2: Confusion matrices by race

BIPOC group, about half of people who *didn't* recidivate were predicted incorrectly, while in the other POC and white groups, about half or over half of people who *did* recidivate were predicted incorrectly.

3.1 Accuracy

Accuracy is a measure of how often the algorithm is correct. Accuracy is defined by

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Let's calculate the accuracy of the overall data set.

$$ACC = \frac{2035 + 2681}{2035 + 2681 + 1282 + 1216}$$

$$ACC = 0.6537289$$

Table 3.1: Accuracy by race

Race	True Negative	True Positive	False Negative	False Positive	Accuracy
All	2681	2035	1216	1282	0.6537289
BIPOC	995	1378	533	808	0.6389338
Other	547	152	222	125	0.6682600
POC					
White	1139	505	461	349	0.6699267

As we can see in Table 3.1, overall, the algorithm is about 65% accurate, while accuracy for BIPOC is closer to 64%, and accuracy for other racial groups is closer to 67%. These numbers, while showing possible bias against BIPOC individuals, are not wildly different, leading many who look simply at accuracy to believe that the COMPAS algorithm is minimally biased.

3.2 Error Rates

In addition to accuracy, we can calculate different error rates. These values show how likely you are to be predicted incorrectly.

3.2.1 False Omission Rate

False Omission Rate (FOR) is a measure of how likely someone is to recidivate if they are predicted to not recidivate.

$$FOR = \frac{FN}{FN + TN}$$

We can use the formula to calculate the FOR for all individuals

$$FOR = \frac{1216}{1216 + 2681}$$

$$FOR = 0.3120349$$

Table 3.2: False Omission Rate by race

Race	False Omission Rate
All	0.3120349
BIPOC	0.3488220
Other POC	0.2886866
White	0.2881250

We can see in Table 3.2 that overall, about 31% of people, 34% of BIPOC individuals, and

29% of both other POC and white people are incorrectly predicted to not recidivate. This means that if a BIPOC individual is predicted to not recidivate, they are more likely than other groups to recidivate within the next two years.

3.2.2 False Discovery Rate

False Discovery Rate (FDR) is a measure of how often people who are predicted to recidivate do not go on to recidivate.

$$FDR = \frac{FP}{FP + TP}$$

We can use this formula to calculate the FDR of the overall data set.

$$FDR = \frac{1282}{1282 + 2035}$$

$$FDR = 0.3864938$$

Table 3.3: False Discovery Rate by race

Race	False Discovery Rate
All	0.3864938
BIPOC	0.3696249

Race	False Discovery Rate
Other POC	0.4512635
White	0.4086651

We can see the FDR for all racial groups in Table 3.3. Here, we can see that BIPOC individuals are incorrectly predicted to recidivate about 37% of the time, while other POC and white individuals are incorrectly predicted to recidivate 45% and 41%, respectively. This means that other POC are most likely to be predicted to recidivate, and then not recidivate, while BIPOC are least likely.

I am using these probabilities to study the bias in this algorithm. I expected the algorithm to be biased against BIPOC individuals, generally the most discriminated against group in the United States justice system. However, these numbers show the algorithm as being biased towards BIPOC individuals. As an individual going into the justice system, the thing that would be most important for you to avoid would be to be predicted to recidivate, face higher bail or harsher punishments, and then carry on to not recidivate, or to have a False Positive result. However, the FDR for BIPOC individuals is actually lower than any other group, meaning that if a BIPOC individual is predicted to recidivate, they are less likely than an individual from another group to not recidivate. These results are the opposite of what I expected. Figure 2.1 seems to show bias, with many more BIPOC individuals are marked as higher risk for recidivism. I can also see in Figure 3.2 that the BIPOC group has many more FP individuals than any other racial group. All of this is because of what we see in Table 2.2,

where we see that BIPOC individuals are more likely in general to be rearrested within two years. So to see how much of these differences are due to machine learning bias, and how much is because of a larger population of BIPOC individuals being rearrested, I wanted to analyze other probabilities, and see if they revealed any bias in the COMPAS recidivism algorithm.

3.2.3 Miss Rate/False Negative Rate

The Miss Rate or False Negative Rate (FNR) is a measure of how often people that recidivated were incorrectly predicted not to recidivate.

$$FNR = \frac{FN}{FN + TP}$$

We can calculate the FNR of all individuals below.

$$FNR = \frac{1216}{1216 + 2035}$$

$$FNR = 0.3740388$$

Table 3.4: False Negative Rate by race

Race	False Negative Rate
All	0.3740388
BIPOC	0.2789116

Race	False Negative Rate
Other POC	0.5935829
White	0.4772257

As we can see, BIPOC individuals who were rearrested within 2 years were significantly less likely to be incorrectly predicted to not recidivate, with 27% being incorrectly predicted, compared to 60% of other POC and 48% of white individuals. That means that while almost half of all white individuals who recidivated, and over half of all other POC individuals who recidivated were predicted incorrectly, just over a quarter of BIPOC individuals were. Because the algorithm was more likely to predict that BIPOC individuals will recidivate, and BIPOC individuals are more likely to recidivate, a smaller percent of those who recidivate are predicted incorrectly, while a larger number of individuals of other races who recidivate are predicted incorrectly.

3.2.4 Fall-Out/False Positive Rate

The Fall-Out or False Positive Rate (FPR) is a measure of how likely someone who did not recidivate was to be incorrectly predicted to recidivate, using the formula below:

$$FPR = \frac{FP}{FP + TN}$$

We can use this formula to calculate the FPR for all individuals.

$$FPR = \frac{1282}{1282 + 2681}$$

$$FPR = 5245$$

Table 3.5: False Positive Rate by race

Race	False Positive Rate
All	0.3234923
BIPOC	0.4481420
Other POC	0.1860119
White	0.2345430

We can see that BIPOC individuals who did not recidivate are much more likely than other racial groups to be incorrectly predicted to recidivate. About 45% of BIPOC individuals who did not recidivate were incorrectly predicted to recidivate, while only 19% of other POC and 23% of white individuals who did not recidivate were incorrectly predicted.

Using FOR and FDR, the COMPAS recidivism algorithm seems biased towards BIPOC individuals and mostly against other POC. Looking at FPR and FNR, it seems strongly biased against BIPOC individuals. To further investigate this bias towards or against BIPOC indi-

viduals, I will be examining some statistical non-discrimination criteria.

4 Statistical Non-Discrimination Criteria

Barocas, Hardt, and Narayanan (2019) defines three statistical non-discrimination criteria. They use the joint distribution of different random variables to decide whether or not an algorithm is fair between groups. In the below criterion, A is the sensitive attribute, in this case race. P(A=a) is the probability that a randomly selected individual is BIPOC, P(A=b) is the probability that the individual is another POC, and P(A=c) is the probability that the individual is white. R is the score, in this case whether they were predicted to recidivate (R=1), or predicted to not recidivate (R=0). Y is the target variable, what actually occurred. P(Y=1) is the probability that a randomly selected individual did recidivate, while P(Y=0) is the probability that they did not recidivate. The equations for separation and sufficiency are originally defined with $P(A \mid B, C)$ instead of $P(A \mid B \cap C)$, but they are equivalent in this case, so I have converted them for succinctness.

4.1 Independence

Independence is meant to measure if the score (R) is affected by race. I could find if R is statistically independent from A, but that would give me a simple true or false answer. Barocas, Hardt, and Narayanan (2019) also give us the below formula, which will give me more insight than a simple "yes" or "no."

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b) = P(R = 1 \mid A = c)$$

Using definition

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

we can find:

$$\frac{P(R = 1 \cap A = a)}{P(A = a)} \stackrel{?}{=} \frac{P(R = 1 \cap A = b)}{P(A = b)} \stackrel{?}{=} \frac{P(R = 1 \cap A = c)}{P(A = c)}$$

$$\frac{0.3030219}{0.5148323} \stackrel{?}{=} \frac{0.0383976}{0.1449958} \stackrel{?}{=} \frac{0.1183809}{0.3401719}$$

$$0.5885837 \neq 0.2648184 \neq 0.3480033$$

Clearly, race and score are not independent. BIPOC individuals are far more likely to be predicted to recidivate. If we look at Table 2.2, we can see that the respective probabilities of the different racial groups are similar to their actual recidivism rate. BIPOC are most likely to be predicted to recidivate and most likely to actually recidivate, other POC are least likely,

and white individuals are in the middle, but much closer to the low rates of other POC than the high rates of BIPOC. However, it is important to note that the numbers above are more extreme than the numbers in Table 2.2. BIPOC are more likely to be predicted to recidivate than they are to actually recidivate, while other POC and white individuals are less likely to be predicted to recidivate than they are to actually not recidivate.

4.2 Separation

Separation is meant to measure how the score is affected by race, while taking into consideration the actual likelihood of the racial groups to recidivate. The first part of measuring separation, measuring if the score is affected by race given that the individual actually recidivated, is defined by:

$$P(R = 1 \mid Y = 1 \cap A = a) = P(R = 1 \mid Y = 1 \cap A = b) = P(R = 1 \mid Y = 1 \cap A = c)$$

$$\frac{P(R = 1 \cap Y = 1 \cap A = a)}{P(Y = 1 \cap A = a)} \stackrel{?}{=} \frac{P(R = 1 \cap Y = 1 \cap A = b)}{P(Y = 1 \cap A = b)} \stackrel{?}{=} \frac{P(R = 1 \cap Y = 1 \cap A = c)}{P(Y = 1 \cap A = c)}$$

$$\frac{0.1910175}{0.2649016} \stackrel{?}{=} \frac{0.0210701}{0.0518436} \stackrel{?}{=} \frac{0.0700028}{0.1339063}$$

$$0.7210884 \neq 0.4064171 \neq 0.5227743$$

This is a measurement of how likely people who actually recidivated were to be predicted to recidivate. This is the complementary probability to the FNR. Just like we saw in Table 3.4,

BIPOC individuals are significantly more likely to be correctly predicted to recidivate.

The second part of separation, measuring if the score is affected by race given that the individual did not recidivate, is defined by:

$$P(R = 1 \mid Y = 0 \cap A = a) = P(R = 1 \mid Y = 0 \cap A = b) = P(R = 1 \mid Y = 0 \cap A = c)$$

$$\frac{P(R = 1 \cap Y = 0 \cap A = a)}{P(Y = 0 \cap A = a)} \stackrel{?}{=} \frac{P(R = 1 \cap Y = 0 \cap A = b)}{P(Y = 0 \cap A = b)} \stackrel{?}{=} \frac{P(R = 1 \cap Y = 0 \cap A = c)}{P(Y = 0 \cap A = c)}$$

$$\frac{0.1120044}{0.2499307} \stackrel{?}{=} \frac{0.0173274}{0.0931522} \stackrel{?}{=} \frac{0.0483782}{0.2062656}$$

$$0.448142 \neq 0.1860119 \neq 0.234543$$

This separation equation measures how likely someone who did not recidivate was to be incorrectly predicted to recidivate. This is the same as the FPR, and we can see that we have the same numbers as in Table 3.5. We see that BIPOC individuals who did not recidivate are by far the most likely to be incorrectly predicted to recidivate. This shows bias in the algorithm, with almost 45% of BIPOC individuals who did not recidivate incorrectly predicted to recidivate, and being subjected to longer jail stays and higher bails.

4.3 Sufficiency

Sufficiency is meant to measure if the error rates of an algorithm are affected by race. The first part of this, measuring how likely someone is to recidivate if they are predicted not to

recidivate, is defined by:

$$P(Y = 1|R = r \cap A = a) = P(Y = 1|R = r \cap A = b) = P(Y = 1|R = r \cap A = c)$$

$$\frac{P(Y = 1 \cap R = 0 \cap A = a)}{P(R = 0 \cap A = a)} \stackrel{?}{=} \frac{P(Y = 1 \cap R = 0 \cap A = b)}{P(R = 0 \cap A = b)} \stackrel{?}{=} \frac{P(Y = 1 \cap R = 0 \cap A = c)}{P(R = 0 \cap A = c)}$$

$$\frac{0.0738841}{0.2118104} \stackrel{?}{=} \frac{0.0307735}{0.1065983} \stackrel{?}{=} \frac{0.0639035}{0.221791}$$

$$0.348822 \neq 0.2886866 \approx 0.288125$$

This equation is measuring the same thing that FOR is (Table 3.2). We see here that the likelihood of other POC and white individuals being incorrectly predicted to not recidivate is about equal, but BIPOC individuals are much more likely to be incorrectly predicted to not recidivate because they are always more likely to recidivate, no matter their score.

The second part of sufficiency, measuring how likely someone is to recidivate if they are predicted to recidivate, is defined by:

$$\frac{P(Y = 1 \cap R = 1 \cap A = a)}{P(R = 1 \cap A = a)} \stackrel{?}{=} \frac{P(Y = 1 \cap R = 1 \cap A = b)}{P(R = 1 \cap A = b)} \stackrel{?}{=} \frac{P(Y = 1 \cap R = 1 \cap A = c)}{P(R = 1 \cap A = c)}$$

$$\frac{0.1910175}{0.3030219} \stackrel{?}{=} \frac{0.0210701}{0.0383976} \stackrel{?}{=} \frac{0.0700028}{0.1183809}$$

$$0.6303751 \neq 0.5487365 \neq 0.5913349$$

This equation is the complementary probability to FDR (Table 3.3). Again, we can see that BIPOC individuals are more likely to be accurately predicted to recidivate, because BIPOC individuals recidivate more overall.

Once again, we see that BIPOC are far more likely to be predicted to recidivate, and far more likely to actually recidivate. We can draw many of the same conclusions from these criteria as we can from our confusion matrix calculations. BIPOC individuals are more likely to recidivate than other racial groups with the same binary score, but BIPOC individuals are also more likely to be incorrectly predicted to recidivate, even if they do not go on to recidivate.

5 Conclusion

In this analysis, I found that more BIPOC people are predicted to recidivate, and almost half of BIPOC people who did not recidivate were incorrectly predicted to recidivate. About half of other individuals who did recidivate were incorrectly predicted to not recidivate. Looking at these numbers, it seems very clear that the COMPAS algorithm is biased against BIPOC individuals.

Yet a BIPOC individual predicted to recidivate is more likely to recidivate than another POC or a white individual predicted to recidivate. This seems to show the COMPAS algorithm being biased towards BIPOC individuals.

So which is more accurate? That depends on your point of view. An official wanting to implement this algorithm could see that accuracy is about the same for all races, FOR is higher for BIPOC people, and FDR is lower for BIPOC people, and conclude that this algorithm is not significantly biased against BIPOC individuals. A BIPOC defendant going in to the justice system could see that FDR is lower for BIPOC people than for other racial groups, and conclude that it is not biased. But a BIPOC individual who has been through the justice

system, did not recidivate, was incorrectly predicted to recidivate, and sees that almost half of all other BIPOC people who did not recidivate went through the same, could conclude that this algorithm is biased against BIPOC individuals.

Whatever you conclude about this algorithm's bias, I believe it shows deeper issues within the American justice system. We see in Table 2.3 that more BIPOC individuals are arrested in Broward County than all other individuals combined. We see in Table 2.2 that BIPOC individuals are over 10% more likely to recidivate than white individuals. Even though the algorithm is marking hundreds more BIPOC as high risk, this doesn't decrease its accuracy, because many of those BIPOC individuals go on to be rearrested. We cannot know from this data set why so many more BIPOC people are arrested and rearrested than other racial groups. It could be because of overpolicing, poverty, stereotypes, or countless other reasons.

This lack of information is one of the weaknesses of my analysis. I only have access to the data of the people who were arrested, not the data about people who committed initial or subsequent crimes and were never caught. BIPOC communities are overpoliced, and many criminals of all races commit crimes without being arrested. This means that my data set may not be complete in both initial arrests and recidivism, and I can't know if any bias that I did find is due to bias in the COMPAS algorithm, or due to bias baked into the policing system. Another of the weaknesses of my analysis is that I did not control for gender, age, or any other confounding variables, so there are many areas of possible bias and pieces of information that I did not analyze. One strength of my analysis is that I have a large data set. I have over 7,000 individuals in the data set, so my analysis is not as strongly affected by outliers or recording

mistakes. Another strength is that my analysis is consistent. From the very basics of looking at the pure number of individuals arrested and rearrested and the number marked as high risk to calculating probabilities, the fact that BIPOC individuals are both arrested more and marked as higher risk shows up over and over in my analysis.

However, this one fact that keeps showing up over and over is frustrating because I cannot reach a clear conclusion on whether the algorithm is biased or not. I went into this exploration expecting to find that the algorithm is unequivocally biased against BIPOC people because of what I have learned about the American justice system. However, I found that in many cases, calculations did not show that. In some instances, like looking at the number of BIPOC individuals marked high risk or looking at the FPR, the algorithm seems biased against BIPOC individuals. Yet when looking at the FOR and other probabilities, the algorithm seems biased towards BIPOC individuals. This exploration helped me gain a deeper understanding of machine learning bias. I learned that machine learning bias really is based off of the bias already inherent in our systems and societies, so it is difficult to diagnose and solve machine learning bias.

So how do we diagnose and solve any possible biases in this algorithm, and others like it? I want to study computer science in college, and one day have a career in it, so my generation of computer scientists will be grappling with machine learning bias, and trying to find solutions. No AI is perfect, so we have to decide what is important to us. Do we want to focus on equalizing error rates? Or do we want to see our own values reflected back at us in our models? The COMPAS algorithm is not wrong in marking many BIPOC individuals high

risk, but would it be better if it marked fewer BIPOC individuals high risk, giving them more chances to be integrated into society and less bail to pay? In Figure 1.1, we see a translation model aggravating gender stereotypes. While this is shocking and angering, it is not incorrect. The medical field, writing, science, and computer programming are all male-dominated fields. So do we want our models to always be the most statistically correct at the expense of BIPOC individuals hoping to turn their life around after one crime, or little girls dreaming of becoming the next big scientist or author or computer programmer? Or do we want to give little girls a chance to dream at the expense of our model? These are choices that my peers and I will have to make, and I for one will now go into these decisions armed with knowledge of how machine bias works, and an understanding of its complexities that I did not have before.

6 References

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

"Fairness Metrics Overview." n.d. *IBM*. https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-fairness-metrics-ovr.html.

Larson, Jeff, Julia Angwin, Lauren Kirchner, and Surya Mattu. 2016. "How We Analyzed the Compas Recidivism Algorithm." *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Pratt, Mary K. 2020. "What Is Machine Learning Bias (AI Bias)?" *Enterprise AI*. TechTarget. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-bias-algorithm-bias-or-AI-bias.

"Recidivism." n.d. *National Institute of Justice*. https://nij.ojp.gov/topics/corrections/recidivism.

Shane, Janelle. 2021. *You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place*. Voracious/Little, Brown; Company.

"What Is Machine Learning?" n.d. *IBM*. https://www.ibm.com/cloud/learn/machine-

learning.