Interpretability and Fairness in Machine Learning

Aylin Jamali 5367428

1. Introduction to Fairness

Machine Learning and especially deep learning approaches are black boxes for us. We give them some data and they give us the results. The seminar “Interpretability and Fairness in Machine Learning” was about two major topics as it is obvious from the name. The first part was about interpretability and in the second part we talked about fairness.
Interpretability is about being able to know how a machine is making its decisions. In this way we can trust the algorithm. Especially in sensitive cases such as healthcare it is crucial to know what exactly the method is doing. It is much easier to debug the model and the implementation if our approach is interpretable. It would be a huge help for data scientists and engineers to improve the models if they are provided insights of the functionality of the model and where it might be going wrong.
On the other hand, Fairness in Machine Learning verifies that the results are not biased or there is no discrimination against individuals or groups based on their race, age, gender, etc. This is essential from the perspective of ethics when we want to use machine learning systems in the daily lives.
In order to remove bias, we should first find the source of the bias. There can be different bias sources: Data, Algorithm, and User experiences.

1. **Data:** Most of the AI and ML algorithms are data driven and require data to be trained. Therefore, data is tightly coupled to the functionality of the algorithms and systems. When you have bias in the data and when it is used by ML training algorithms, it might result in biased algorithmic outcomes. There can be measurement bias, omitted important variable bias, representation bias, etc. in our data.
2. **Algorithm:** On the other hand, there might be bias in the algorithm itself. Or you can have a popularity bias. Items that are more popular tend to be exposed more. As an example, search engines or recommendation systems might present these popular objects more to the public.
3. **User Experience:** The last source can be user experience bias. Behavioral and Social biases can be two examples. Social bias happens when others’ actions affect our judgment.

2.1 Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments

The paper “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments” studies fairness in the recidivism prediction instruments (RPIs). These devices provide an assessment of likelihood that a criminal defendant will reoffend at a future point in time. Risk assessment instruments are getting more popular within criminal justice systems. They are categorized as high-risk classifications since it affects the defendants lives. One of the known RPIs is COMPAS, developed by Northpointe, Inc. There has been an investigation on this device by a team at ProPublica, Angwin et al. They concluded that COMPAS is biased against African-American, here categorized as black, defendants. They claim that the likelihood of a non-recidivating black defendant being assessed as high risk is nearly twice that of white (the Caucasian) defendants. Similarly, the likelihood of a recidivating black defendant being

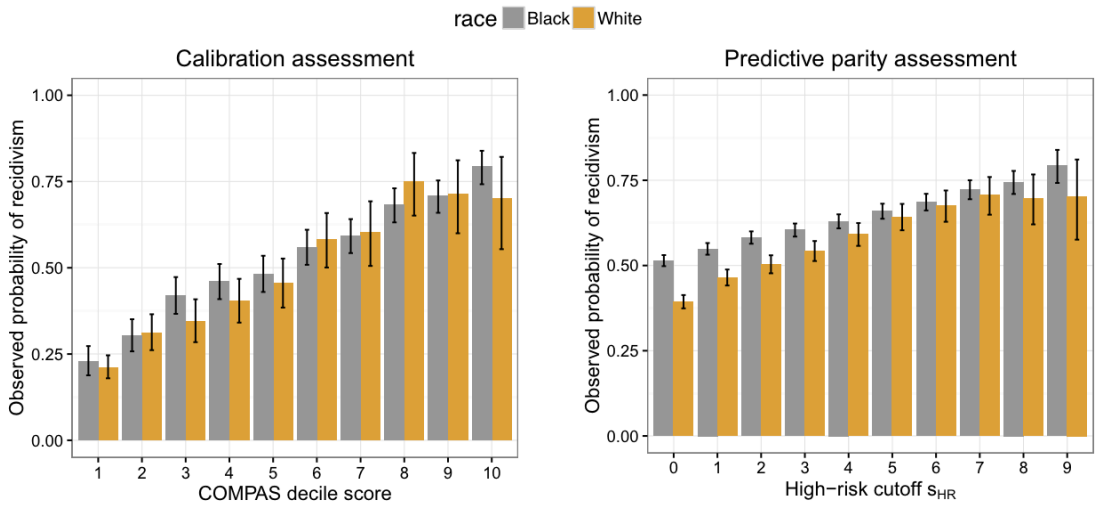
assessed as low risk is nearly half that of white defendants. In other words, COMPAS has higher False Positive Rates (FPR) and lower False Negative Rates (FNR) for black defendants than for white defendants. However, Flores et al. argues that COMPAS satisfies the calibration, which is another fairness criterion, and the correct approach to assess an RPI is to evaluate this criterion instead of FPR and FNR. This device also satisfies another fairness criterion, predictive parity.

Before we dive into the paper it is important to know about these criteria and their definitions. If we want to design models with less or in the best case no bias, or if we want to evaluate the fairness of RPIs, we have to understand these criteria. As a background, let $S = S(x)$ denote the risk score based on the covariates $X = x \in R^p$, and let $R \in \{b, w\}$ denote the group to which the individual belongs. Group b refers to defendants with race African-American and group w refers to Caucasian people. Also in $Y \in \{0, 1\}$, when $Y = 1$ it means that the defendant goes on to recidivate. The quantity s_{HR} denotes the high-risk score threshold. Now we want to introduce the fairness criteria discussed in the paper:

1. **Calibration:** A score $S = S(x)$ is said to be well calibrated if it reflects the same likelihood of recidivism irrespective of the individuals' group membership. That is, if for all values of s,
   $$P(Y = 1|S = s, R = b) = P(Y = 1|S = s, R = w). \quad \text{(Eq. 1)}$$
   In the context, it means that if the predicted risk scores can show the actual likelihood in high accuracy, we call the system well-calibrated. We can say that our model is calibrated very well when the risk scores it assigns are consistent with the actual results for the groups b and w. For example, COMPAS is known for being well-calibrated. It means that its risk scores are equally reliable and meaningful in various racial groups.
2. **Predictive Parity:** A score $S = S(x)$ satisfies predictive parity at a threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same regardless of group membership. That is, if
   $$P(Y = 1|S > s_{HR}, R = b) = P(Y = 1|S > s_{HR}, R = w). \quad \text{(Eq. 2)}$$
   Predictive parity makes sure that the accuracy of positive predictions (when someone is predicted that will reoffend) is the same for different racial groups. So, if the tool predicts people from different groups will reoffend, it should be equally correct for all groups. ProPublica analysis shows that COMPAS satisfies predictive parity for threshold choice of interest.
3. **Error Rate Balance:** A score $S = S(x)$ satisfies error rate balance at a threshold s_{HR} if the false positive and false negative rates are equal across groups. That is, if
   $$P(S > s_{HR}|Y = 0, R = b) = P(S > s_{HR}|Y = 0, R = w) \quad \text{(Eq. 3)}$$
   and
   $$P(S \leq s_{HR}|Y = 1, R = b) = P(S \leq s_{HR}|Y = 1, R = w). \quad \text{(Eq. 4)}$$
   If you ensure that the error rate is balanced, it means that the system makes sure that no group is harmed more by wrong predictions.

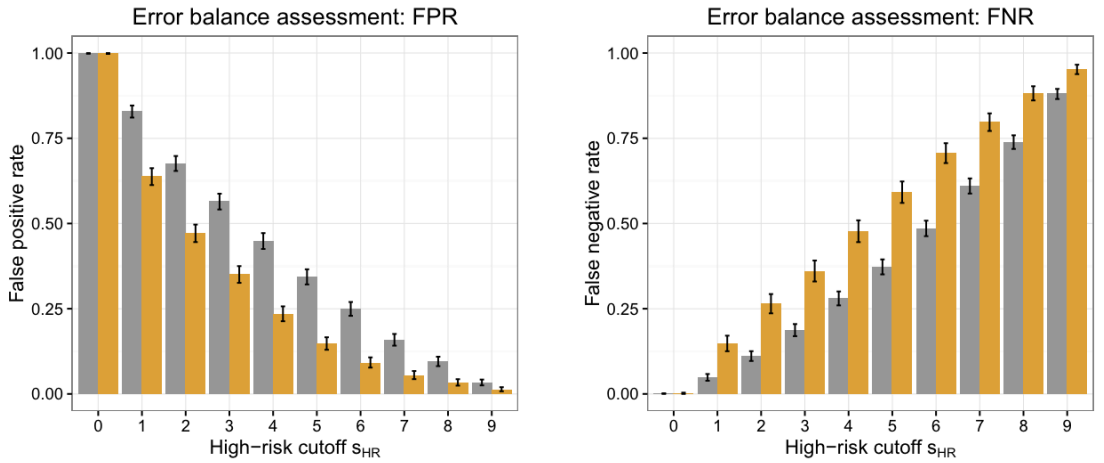
This article first highlights the complexities and trade-offs involved in meeting these fairness criteria simultaneously. For instance, satisfying error rate balance may lead to disparities in predictive parity, and vice versa. Secondly, they show how a risk prediction tool with different error rates across groups can unfairly affect certain groups when stricter penalties are given to

those deemed high-risk. They use "disparate impact" to mean when a policy unintentionally harms one group more than others.



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

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



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

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

Figure 1: Empirical assessment of the COMPAS RPI according to three of the fairness criteria presented in Section 2.1. Error bars represent 95% confidence intervals. These Figures confirm that COMPAS is (approximately) well-calibrated, satisfies predictive parity for high-risk cutoff values of 4 or higher, but fails to have error rate balance.

2.2 Predictive parity, false positive rates, and false negative rates

The first main result of this paper is to show that predictive parity cannot coexist with error rate balance when the prevalence of outcomes varies between groups. In Figure 1 you can see the observed recidivism rates and error rates corresponding to fairness criteria calibration, predictive parity, and error rate balance. As you can see, this RPI is well calibrated and satisfies

the predictive parity for high-risk threshold $s_{HR} > 4$. However, it fails to satisfy the error rate balance (false positive rate and false negative rate) across the range of high-risk cutoffs. Even if you increase the threshold from 4 to 7, this problem persists. Here they showed that the error rate imbalance is not coincidence and if the recidivism prevalence is different across groups and an RPI satisfies predictive parity at a given threshold, it must have imbalance FPRs and FNRs at the threshold. In other words, the predictive parity and error rate balance are mutually exclusive when recidivism prevalence is different.

As we know from the predictive parity definition, predictive parity means that when we set a specific risk threshold, the accuracy of positive predictive value (PPV) should be the same for all groups. We can relate prevalence (p), PPV, and FPR and FNR in this equation:

$$FPR = \frac{p}{1-p}\frac{1-PPV}{PPV}(1 - FNR). \quad \text{(Eq. 5)}$$

Then it can be understood that if the predictive parity is satisfied (means that if the PPV is the same across groups) but the prevalence (p) is different between groups, we cannot have the same FPRs and FNRs across the groups.

Here in the COMPAS data, we see that recidivism among black defendants is 51% and among white defendants is 39%. Therefore with any threshold, we will have an imbalance between false positive and false negative rates if an RPI satisfies predictive parity.

2.3 Assessing impact

Now that we know that not all fairness criteria can be satisfied at the same time, it is important to assess the impact of imbalance in FPRs and FNRs where we have the policy of high-risk assessment resulting in a stricter penalty. The risk assessments can be used to announce bail, parole, or sentencing decisions. In this section they analyzed use cases that high risk individuals receive stricter penalties. It is good to know that in general there exists some cases where high risk individuals receive benefit rather than a penalty.

Here they used a simple risk-based MinMax policy, that the defendant receives a penalty $t_{min} \leq T \leq t_{max}$:

$$T_{\text{MinMax}}(s) = \begin{cases} t_{\min} & \text{if } s > s_{\text{HR}} \\ t_{\max} & \text{if } s < s_{\text{HR}} \end{cases} \quad \text{(Eq. 6)}$$

I think there has been a mistake in this paper. They specifically mention that “the analysis here addresses use cases wherein high-risk individuals receive stricter penalties”. Then in their MinMax policy they give t_{min} penalty if $s > s_{HR}$ and t_{max} penalty if $s < s_{HR}$. This means that if the risk is more than the threshold, they assign smaller (t_{min}) penalty and if the risk score is less than the threshold, they assign bigger (t_{max}) penalty which contradicts their own claim mentioned above.

By ignoring this mistake, we continue by defining the disparate impact with the difference between the sentence duration of defendants in different racial groups with different outcomes:

$$\Delta = \Delta(y_1, y_2) \equiv \mathrm{E}(T|R = b, Y = y_1) - \mathrm{E}(T|R = w, Y = y_2). \quad \text{(Eq. 7)}$$

If we use the MinMax policy to calculate the difference in the penalties we get:

$$\Delta \equiv \mathrm{E}(T|R = b, Y = y_1) - \mathrm{E}(T|R = w, Y = y_2)$$
$$= (t_{max} - t_{min})(P(S > s_{HR}|R = b, Y = y_1)$$

$$- P(S > s_{HR} | R = w, Y = y_2)) \quad \text{(Eq. 8)}$$

We can deduct two corollary:

1. For nonrecidivists the difference in average penalty under the MinMax policy is:

$$\Delta = (t_{max} - t_{min})(FPR_b - FPR_w) \quad \text{(Eq. 9)}$$

2. For recidivists, the difference in average penalty under the MinMax policy is:

$$\Delta = (t_{max} - t_{min})(FNR_w - FNR_b) \quad \text{(Eq. 10)}$$

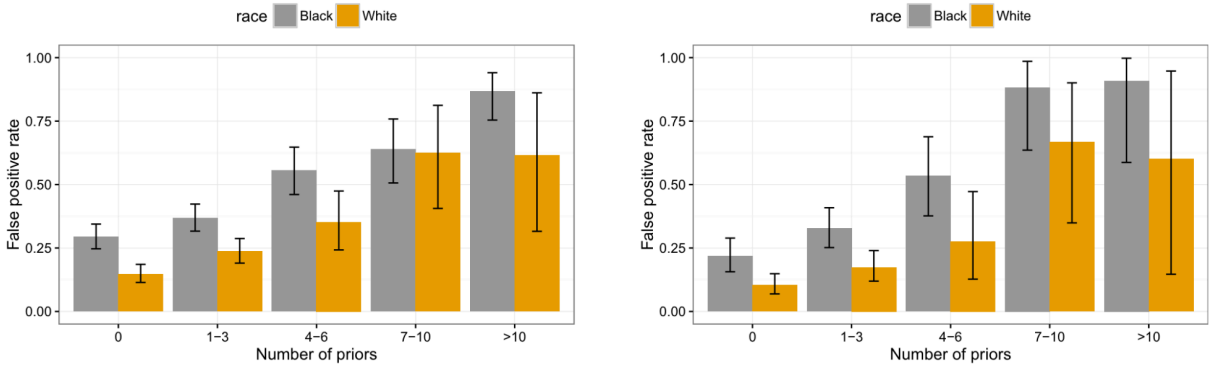
So if an RPI satisfies predictive parity and is used for groups with different recidivism prevalence, the higher recidivism prevalence group will have higher FPR and lower FNR. From both corollaries we can deduct that on average this will result in greater penalties for the group with higher recidivism prevalence, no matter if they are recidivists or nonrecidivists.

There is a special case related to the sentencing decisions where $t_{min} = 0$. This situation might occur for offenders convicted of low-severity crimes who have good prior records. Instead of incarceration, offenders in this category might receive restorative sanctions (alternatives to jail time). If then $t_{max} = 1$, we can say that the expected sentence $E[T]$ can be interpreted as the probability that a defendant receives any period of incarceration. Where $P(T \neq 0)$ is the probability that the sentence is not zero,

$$E[T] = P(T \neq 0), \quad \text{(Eq. 11)}$$

which means that some period of incarceration is imposed.

Here comes the comparison between the FPR for defendants in group b and w. If someone in group b, who will not reoffend, is $\frac{FPR_b}{FPR_w}$ times more likely to be incarcerated compared to a non-recidivist in group w, this indicates a disparity in error rates between the groups. They examined whether these overall differences in error rates remain if it is looked at more specific subpopulations. They also explored more by analyzing whether the disparities in error rates observed at a broader level also exist within smaller subpopulations.



(a) All defendants.

(b) Defendants charged with a Misdemeanor offense.

Figure 2: False positive rates across prior record count. Plot is based on assessing a defendant as "high-risk" if their COMPAS decile score is $> s_{\text{HR}} = 4$. Error bars represent 95% confidence intervals.

FPRs are higher for defendants with more serious criminal histories in both racial groups. However, significant differences in FPRs between black and white defendants repeat even for

the person with fewer prior offenses. Figure 2 shows that these differences remain even for those charged with misdemeanors, the lowest severity offense category. Additionally, the paper looks at how their fairness results support when they use more detailed features about the defendants. It shows the robustness and the thoroughness of their analysis. They even explore how different punishments might change if they have specific details about the defendants. They also look at how differences in risk scores between black and white offenders can affect their sentences. It links these score variabilities to differences in sentence lengths. They use a measure called total variation distance to measure these gaps better. It works well even if the score data is not normally distributed. They also include a mathematical formula that shows how sentencing differences relate to how much the score distributions overlap.

$$\Delta(y_1, y_2) \leq (t_{max} - t_{min}) d_{TV}(f_{b,y1}, f_{w,y2}) \quad \text{(Eq. 12)}$$

Here $\Delta(y_1, y_2)$ represents the difference in average penalties between two groups, t_{max} and t_{min} are the maximum and minimum sentencing times, and d_{TV} is the total variation distance between the score distributions $f_{b,y1}$ and $f_{w,y2}$ for the two groups.

2.4 Revisiting predictive parity

In the last section they revisited the notion of predictive parity. We know that we have no direct control over recidivism prevalence, but we can have some control over PPV and error rates. We can tune the classifiers in three ways:

1. Allow unequal FNRs to retrain equal PPVs and achieve equal FPRs.
2. Allow unequal FPRs to retrain equal PPVs and achieve equal FNRs.
3. Allow unequal PPVs to achieve equal FPRs and FNRs.

Based on Equation 5, FPR is a linear function of FNR under constraints on PPV and p. If we fix PPV, the first method might need a very large increase in FNR to balance FPR. The same can happen with strategy 2. Because they might reduce disparate impact for one group but increase it in the other group.

The third strategy might achieve by allowing high-risk threshold $s_{HR,r}$ to differ across groups.

2.5 Empirical results

The results of this paper are based on the data collected by Broward County and it is publicly available by ProPublica. It contains COMPAS recidivism risk decile scores, 2-year recidivism results, age, race, sex and some crime-related variables on defendants who were scored in 2013 and 2014. After some data preprocessing, there are $n = 6150$ defendants which $n_b = 3696$ are defined as African-American and $n_w = 2454$ are Caucasian. The results come from Broward County but they used the sentencing guidelines of Pennsylvania State.

Their empirical results are based on two hypothetical sentencing rules:

1. MinMax rule
2. Interpolation rule

The Interpolation rule, unlike the MinMax policy, is not based on a threshold score. It interpolates sentences linearly between t_{min} and t_{max} based on the assigned decile score:

$$T_{Int}(s) = t_{min} + \frac{s-1}{9}(t_{max} - t_{min}). \quad \text{(Eq. 13)}$$

Penalty ranges t_{min} and t_{max} were selected by matching each offender's charge degree (M2 - F1) to sentence ranges in Pennsylvania's Basic Sentencing Matrix. The matrix shows the ranges of the sentences based on two factors: First, how severe the current offense is and second, the defendant's prior record scores. These scores have a range from 0 to 5+. Since the information in the Broward County data was limited, the researchers could not reliably assign prior record scores to the defendants. For their analysis, they used the sentencing range corresponding to a prior record score of 1 for all defendants.
Figure 3 shows the expected sentences for black and white defendants. They are categorized by their observed recidivism result. The x-axis represents the offense gravity score, which was mapped to charge degrees as shown here:

Offense gravity score	2	3	5	7	8
Charge degree	(M2)	(M1)	(F3)	(F2)	(F1)

Table 1. Mapping between offense gravity score and charge degree in the empirical analysis

Based on Figure 3, we can observe that black defendants receive higher sentences than white defendants. They both are in the non-recidivating subgroup and in the recidivating subgroup. This shows that the empirical result is consistent with the theory presented above. White defendants have higher FNRs and lower FPRs than black defendants.

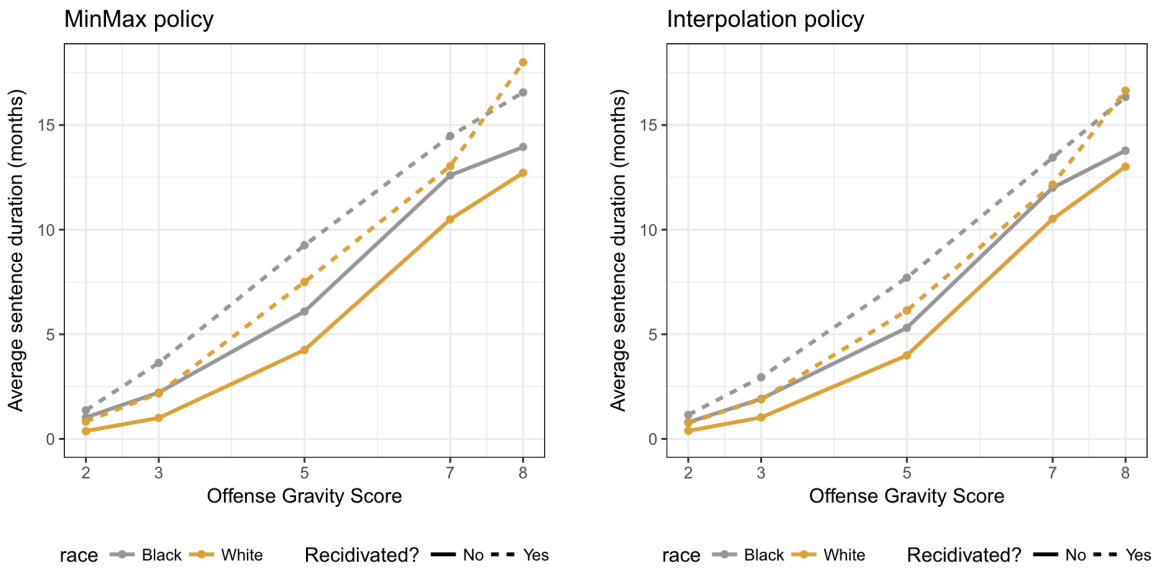


Figure 3. Average sentences under the hypothetical sentencing policies. The mapping between the x-axis variable and the offender's charge degree is given in Table 1. For all OGS levels except 8, observed differences in average sentences are statistically significant at the 0.01 level.

3. Case Study using my methods on the ProPublica dataset

I conducted an empirical analysis of the COMPAS risk assessment tool using the same dataset, focusing on fairness criteria and potential disparate impact. You can find the codes and the figures here.

The analysis included the following key components:

1. Error Rate Analysis: I examined false positive rates (FPR) and false negative rates (FNR) across different risk score thresholds. This analysis revealed significant differences in error rates between racial groups, with African-American defendants experiencing higher FPRs and lower FNRs compared to Caucasian defendants.
   As you can see in Figure 4, significant differences in false positive and false negative rates were observed between black and white defendants. It remains the same across various risk thresholds.
2. Calibration Assessment: I created a calibration plot comparing the COMPAS scores to observed recidivism rates for black and white defendants. Figure 5 helps visualize whether the tool predicts risk equally well for both racial groups. As we can see, it showed reasonable calibration across racial groups, but with some discrepancies at higher risk scores.
3. Predictive Parity Evaluation: It was assessed if the positive predictive value (PPV) was similar across different groups for various risk score thresholds. Figure 6 determines if the tool's predictions are equally accurate for both groups when classifying individuals as high-risk. The data revealed higher average scores for black defendants compared to white ones, even when controlling for recidivism outcome.
4. Disparate Impact Simulation: Last but not least, two hypothetical sentencing policies (MinMax and Interpolation) were simulated based on the scores. Figure 7 demonstrates that the differences in risk scores could lead to disparate impacts, with African-American defendants receiving longer average sentences under both hypothetical policies. However as you can see the results from my analysis are not completely compatible with Figure 4 in the paper. However in general it can be observed that the sentence for African American race is higher than Caucasian group regardless of their recidivation status. Also as it is mentioned in the paper, since the sample size for F1 charge degree is small, it does not follow the general trend and the results are non-significant. So in general the pattern shows that my results also support the theory mentioned in the paper. The general trend is the same, but the sentence numbers are different between my results and the figure included in the paper which might be due to the difference in preprocessing and handling small sample size.

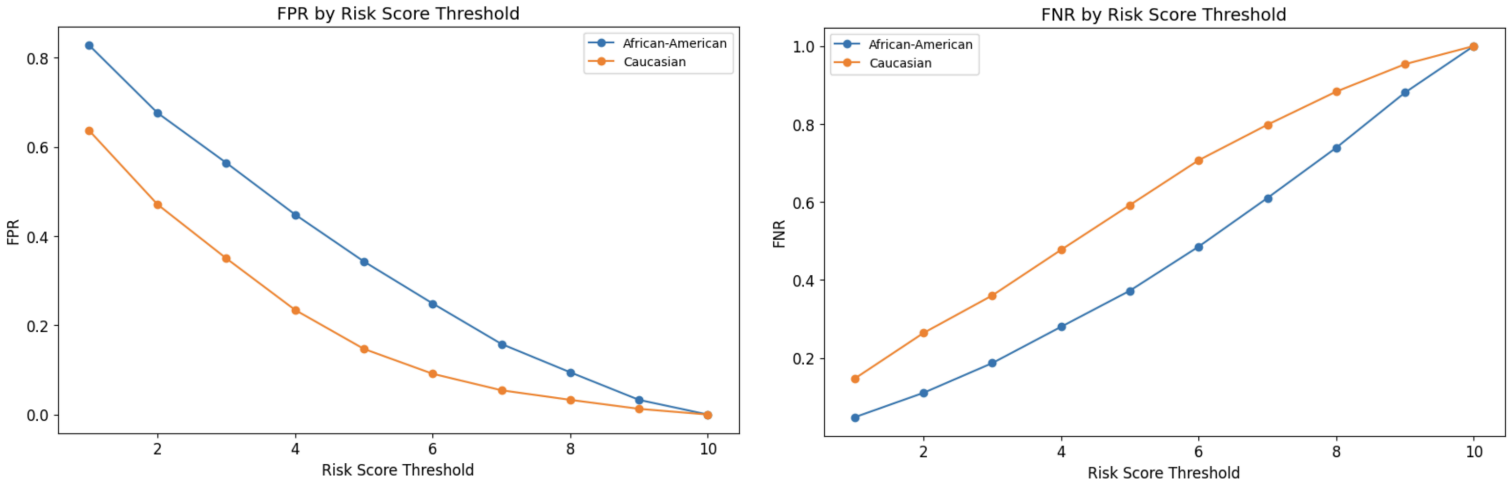


Figure 4. In the x-axis you can find various risk thresholds and in the y-axis false positive rate and false negative rates are shown for both racial groups. FPR for black defendants is higher than white defendants and FNR for black defendants are lower than white ones.

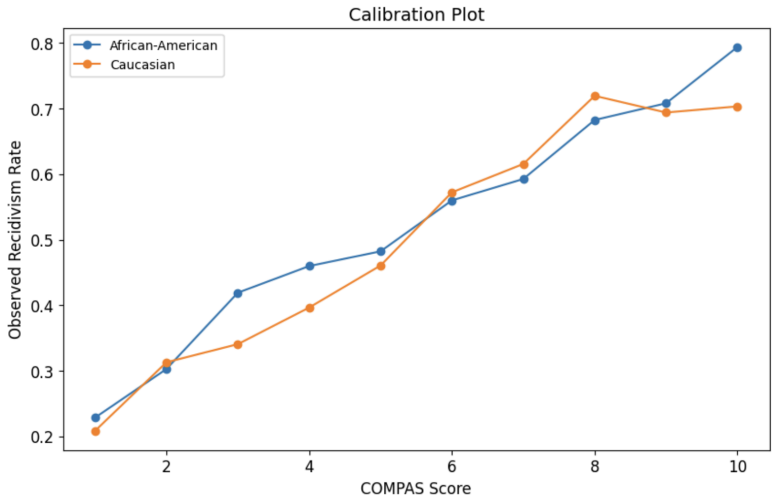


Figure 5. In the x-axis you can see the decile scores (1-10), with higher scores indicating higher predicted risk and the y-axis shows the observed recidivism rate for each decile score. As we can see the lines for both racial groups closely follow each other and increase steadily from left to right which suggests that COMPAS is well-calibrated across races.

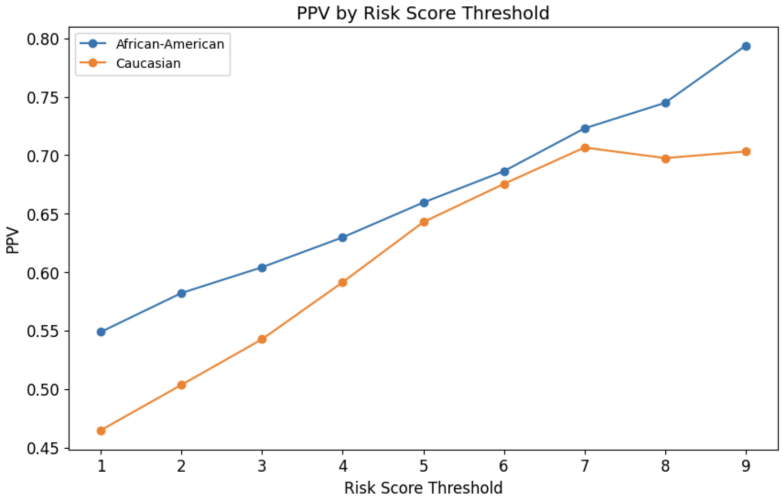


Figure 6. It shows the difference in the positive prevalence value among racial groups across various risk score thresholds.

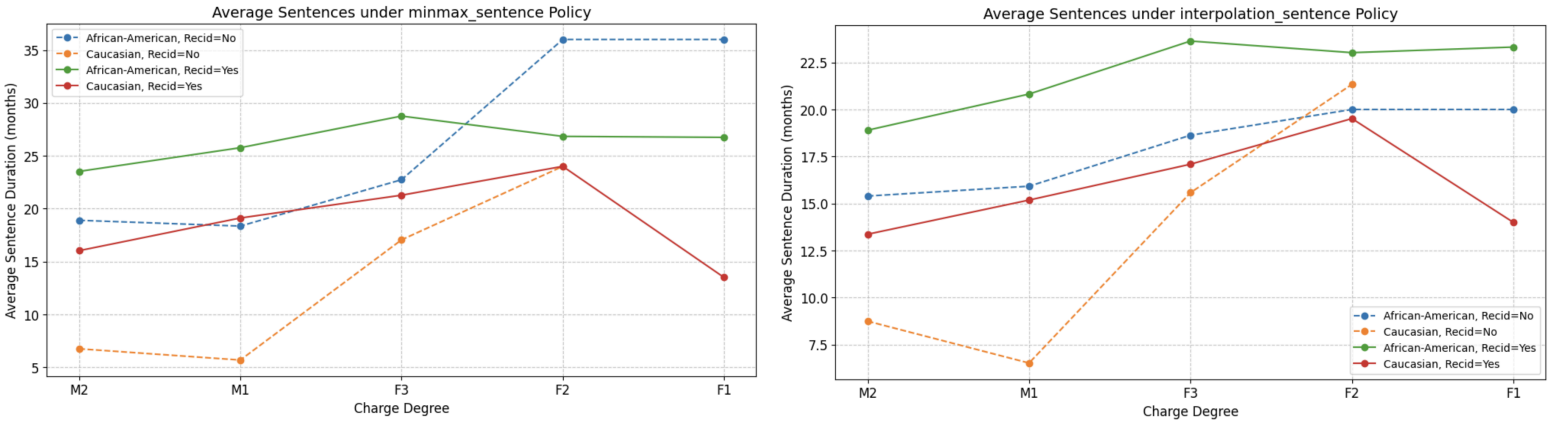


Figure 7. The average sentence under both MinMax and Interpolation policies. Based on the information provided in the paper, the t_{min} and t_{max} values of 0 and 36 months respectively were chosen to approximate the sentencing guidelines used in Pennsylvania. You can find the mapping table between charge degrees and the scores in Table 1.

Furthermore I did some experiments to check whether different sex in the racial group has any impact on the recidivism.

Sample size:

1. African-American: 652 females, 3044 males
2. Caucasian: 567 females, 1887 males

This shows a larger sample size for African-Americans, especially males.

Recidivism rates:

1. African-American: Females 37.9%, Males 54.3%
2. Caucasian: Females 35.1%, Males 40.6%

We observe higher recidivism rates for African-Americans, with a particularly large difference between black and white males.

COMPAS score distribution:

1. African-American: Females mean 4.76, Males mean 5.50
2. Caucasian: Females mean 3.90, Males mean 3.69

African-Americans, especially males, receive higher average COMPAS scores.

False Positive Rates (FPR):

1. African-American: Females 25.2%, Males 21.1%
2. Caucasian: Females 19.6%, Males 12.6%

African-Americans have higher FPRs, with the highest rate for African-American females.

False Negative Rates (FNR):

1. African-American: Females 11.3%, Males 15.0%
2. Caucasian: Females 15.2%, Males 19.9%

Caucasians have higher FNRs, with the highest rate for Caucasian males.

The data and the exploration on the gender differences showed that in the same racial group, males generally have higher recidivism rates and COMPAS risk scores than females. In order to check if the difference is statistically significant, I did chi-square test and you can find the results here:
Chi-square statistic: 150.14
p-value: 2.46e-32
The large chi-square value shows that the patterns we see in the data for race, sex, and recidivism are very different from what we would expect if these factors weren't related to each other at all. Since p-value is extremely small, we can say that these differences are very unlikely to happen randomly. It means the result is highly significant. Although a significant association can interfere, it is important to remember that it does not imply causation. The relationships between race, sex, and recidivism are complex and influenced by many factors.

4. Limitation and Potential Extension of the Study

In this part, I mention some ways to improve or expand the topic. I also look at some limitations of the approach they studied in their paper. In this study, they focus mainly on race, but they do not address other factors like gender, age, or socioeconomic status, which can affect the results

to an extent. We could also experiment how various demographic factors have an impact on assessing risks and biases. Furthermore, the current approach uses a binary classification (high-risk vs. low-risk). Recidivism risk is a very complex topic and a binary classification may oversimplify it. We can use more detailed risk levels or a continuous scale. In order to make our model more robust, we could explore how the findings generalize to other data sources. Moreover, there are other important factors and features that their approach did not take into account. These factors might affect recidivism, such as access to rehabilitation programs, socioeconomic conditions, or the support from the community. Including these factors could make the risk assessment more complete.

5. Conclusion

In conclusion, the study of fairness in machine learning, especially recidivism prediction instruments, includes significant complexities and challenges. The tools like COMPAS is there to help the criminal justice system. However important ethical questions about bias and fairness can be raised. Calibration, predictive parity and error rate balance are criteria for evaluating these tools. But since they have inherent trade-offs it is impossible to achieve all of these criteria at the same time. Based on the analysis, differences in recidivism rates across groups can cause imbalances in false negative and false positive rates. Therefore, it might affect the sentencing results.
Future work in this area can improve these tools to make sure that they do not have bias and their results are fair and effective. This can be done by exploring more into the field of fairness and considering other factors.