Exploring Fairness in a COMPAS data set

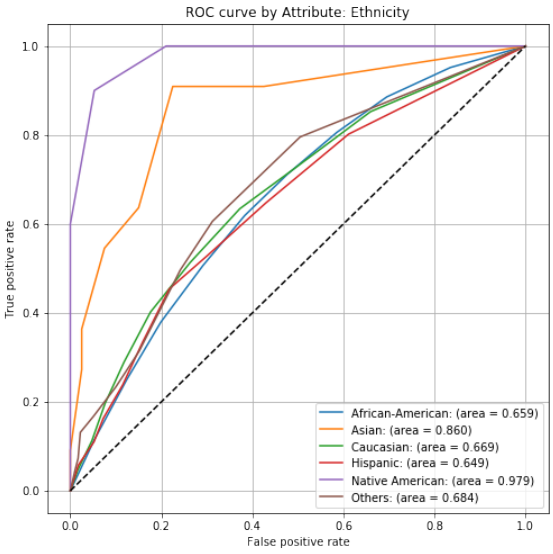
August 30, 2020

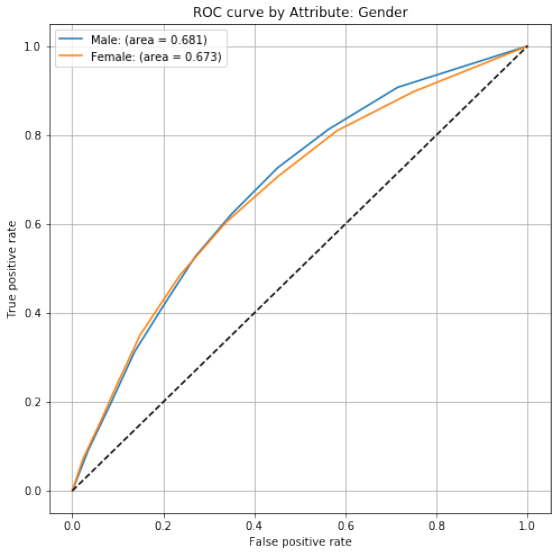
1 Group Project, Maschinelle Intelligenz und Gesellschaft, SoSe20

Laura Lucaj, Maheep Tripathi, Malina Mekhail, Nikolas Hars

1.1 Introduction

The recent events of the Black Lives Matter movement bring into the spotlgiht the fraught issues of racial discrimination in policing. Tech companies that use AI and ML to build policing tools to claim improve fairer and more efficient use of police resources. with products that predict recidivism or probability of occurence of crime using tools such as the COMPAS score. It is very important to make sure AI systems are as fair as possible when used in delicate contexts that determine the future life of an individual. In order to check for fairness in this analysis, we look at a data set of COMPAS recidivism scores and examine it as per various fairness criteria. We select 2 sensitive attributes for our analysis - ethnic background and gender.





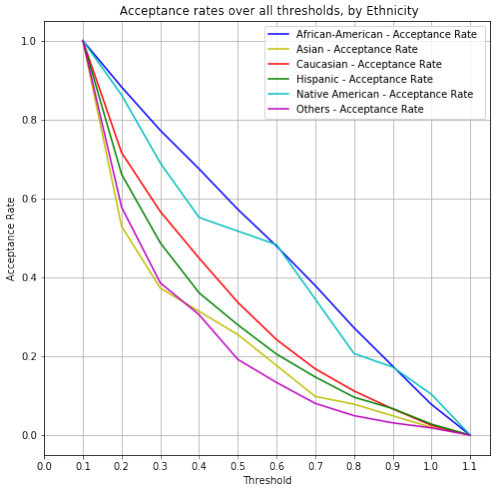
1.1.1 Interpretation of ROC Curves by Attribute

In this plot, we compared the different rates for the african american and caucasian group, as they are the ones with the highest number of samples in the dataset and therefore some significant inferences can be made. Predicting african-americans as more recidivistic, comes with a high cost in civil liberties rights, because a false positive prediction directly impacts the life of the individual, by determining an unfair future based on an inaccurate prediction, which is more severe than negatively predicting a low likelihood of recidivism.

Already we've seen that the classifier performance is not great for ethnic groups. We now examine how well the COMPAS score performs on metrics of fairness - independence, separation and sufficiency.

1.2 TASK 1

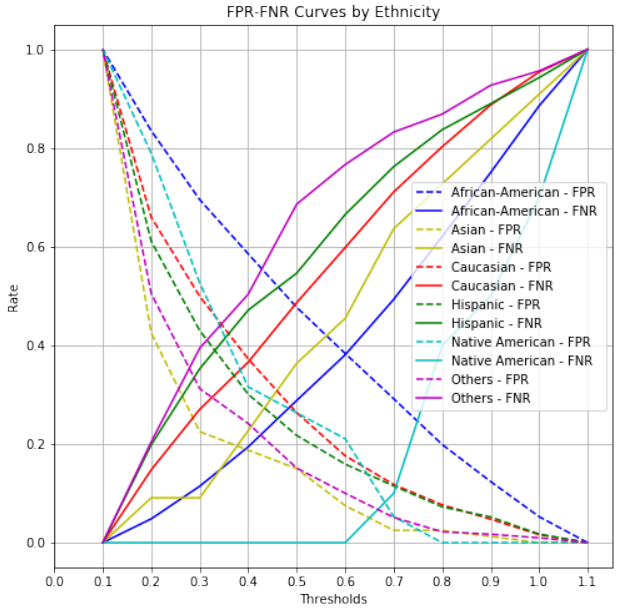
1.2.1 Independence - $R \perp A$



Clearly the COMPAS score is not independent of the sensitive attribute of ethnic background and correlates with the outcome. Looking at the different shapes of the curves it is clear that the acceptance rates are very different for the various ethnical groups. The african-american group, for instance, has the highest acceptance rates across all thresholds and their scores seem to be uniformly distributed in (0,1). This means that this group is more likely to be classified as recidivistic. In comparison, the Caucasian group have a distribution that is skewed towards lower COMPAS scores. A fair classifier should enable everyone to have the same chances, so in this case to make it fairer different thresholds have to be established for the acceptance rates.

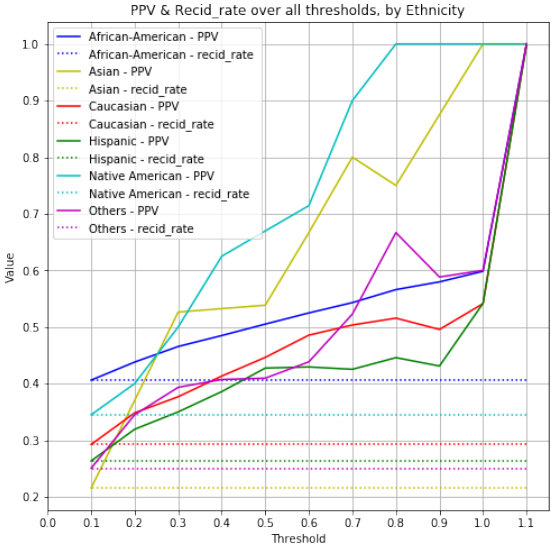
Independence holds only at thresholds 0 and 1 as these are the only thresholds that give the same acceptance rates for all ethnicities. However these are not viable classifiers.

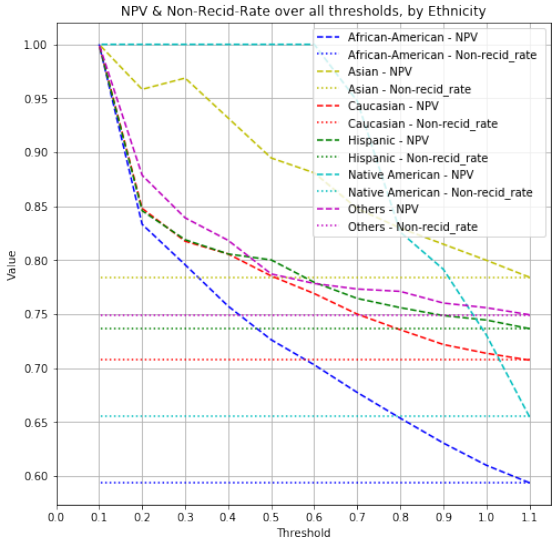
1.3 Separation - $R \; A|Y$



In this plot we can clearly observe that the rates (FPR, FNR) are not equal for both false positives and false negatives. A false positive prediction is more costly, in these circumstances, because the freedom of an individual might directly be harmed by an inaccurate prediction. In this plot, we can observe that the African American group has the highest rate in false positive predictions and the lowest rate in false negative predictions, which again points to the unfairness of the model because the dataset is clearly disproportionately **falsely** targeting this group as more likely to re-offend in the future compared to other groups. When considering that the samples analysed by the model were unequally distributed and consisted of a high-number of African American samples. Ideally a separation-satisfying fair model should have equal false negatives and false positives rates for each group, but this imbalance shows the structural bias present in the american society being mirrored in these algorithms.

1.4 Sufficiency - $Y \perp A|R$





In general we can observe that the classifier does not provide a high dynamic range, which implies that the classifier is not good at all and it is very likely to undertake random predictions. By setting the threshold to high court, high ppv and npv can be reached. However, since the ppv count has a higher dynamic range and it is ethically more important to correctly classify positiv cases, it is better to choose a value where ppv is higher. Unfortunately, the African American group has a ppv range between approx 0,4 and 0,6. From that we can conclude that whatever threshold is chosen, there will be around 50% falsely classified out of all positiv classified. Furthermore, this plot shows that the sample is dispoportionally classifying the African Americans as more recidivistic, and could lead the classifier to determine a high correlation between the African American ethnical background and the likelihood to re-commit a crime in the future. Addtionally, Caucasians also have a small range and an even lower ppv count so there will be even more falsly positively classified people relative to the total amount of positively classified. This also gets obvious by looking at the performance of the predicted value. In general, the model is clearly not satisfying sufficiency because the sensitive characteristics are still determining different npv and ppv.

---

2 TASK 2

The previous section shows that the COMPAS score does not satisfy any of the fairness criteria we examined it for. We attribute this unfairness to the bias in the nature of how and how much training data, per group, is collected. One can expect the nature of the sampled African Americans to be riddled with the structural

racial bias against the group in reporting crime and harsher policing of African Americans. The number of African Americans sampled in this data set is disproportionately higher compared to the size of the group in the population - ideally, more training data would lead to better performance, but in this case it amplifies the error of falsely predicting recidivism among African Americans due to bias. Ideally, we should account for this societal structural bias that show up in the training data, and ensure that ethnicity plays no role in scoring the recidivism probability of an individual. In this task, we try to see if we can achieve any of the fairness criteria by using different thresholds for each group.

2.1 Observations from trying to satisfy independence

Again, ideally the acceptance rates should be equal among the groups, in order for independence to hold. However the implementation of independence used here introduces a relaxation and calculates thresholds for acceptance rates with the minimum distance. Best case, the thresholds should be equal. From this plot we can read that the thresholds need to be different for each group, which indicates that the model is not fair.

	African-American	Asian	Caucasian	Hispanic	Native American	Others
Independence	0.479857	0.514286	0.447466	0.476499	0.500000	0.589610
Seperation: tpr	0.624727	0.833333	0.627841	0.625000	1.000000	0.795918
Seperation: fpr	0.379049	0.448276	0.373767	0.420935	0.250000	0.519164
Sufficiency: ppv	0.534204	0.277778	0.406998	0.357143	0.666667	0.343612
Sufficiency: npv	0.703959	0.941176	0.804623	0.804954	1.000000	0.873418

Theoretically in formal mathematical terms, independence and sufficiency can not both hold together given that the sensitive attribute ethnical background and the outcome (recidivism) are not independent, which is the case in our analysis. These Thresholds look like they are not completely fair but if you consider a soft margin for every metric all of the criteria can nearly hold.

2.2 Observations from trying to satisfy sufficiency and separation tables

From the tables in the notebook we can understand that sufficiency and separation (analysing them separately, not together) do not hold for a hard criteria. However, we found a soft margin in which sufficiency can hold. This margin is the ratio between all PPVs and between NPVs must be greater than 0.75, which means that if PPVs (or NPVs) are up to 25% different from each other, sufficiency will still hold. This points to an unfair model, as they should ideally be equal. Similarly, separation holds only for the trivial cases (thresholds 0.1 and 1.1, but these are not valid classifiers) or if you introduce a soft margin of 75%.

---

3 Task 3 Observations

Accuracy of the Classifier Model: 0.68594

We decided use a decision tree as a binary classifier because we are able to see exactly which of the features are used to classify a defendant as recidivistic and therefore, we can make sure to observe that the sensitive attributes such as ethnical background or gender are not determining the prediction of the likelihood to re-commit a crime in the future. We modified the tree many times in order to make our model as fair as possible, as we wanted to exclude the sensitive attributes of gender and ethnic background for being a discriminative attribute in the decision-making context of the algorithm. the tree seems to be fair for ethnic background as the likelihood of recidivism of the defendants doesn't depend on this attribute.

We modified the code many times, by defining different random states that would give us results that tried to exclude gender and ethnic background in the predictions. Before it gets deployed in such contexts, fairness must be improved and assured, which is mainly depending on the data samples that are fed to the algorithm.

In conclusion it was very hard to modify the code in order to achieve a fairer classifier and based on the fact that we changed the random state in order to achieve a fair model this still does not mean that it is easy to

reproduce and therefore it is not a reliable classifier in such delicate context. The ROC curve below shows that all ethnic groups have the same AUC (Area Under the Curve), which means that the ethnic background is not a decisive factor in the predictions about the individual's likelihood to re-commit a crime.