CS 6603: AI, Ethics, and Society
Written Critique: What-If Tool

Cleo Zhang
yzhang3761@gatech.edu

1 DEFINITION FOR THE TERMS

False Negative – means that a positive entity is misclassified as negative.

False Positive – means an entity that is negative but misclassified as positive.

Accuracy – is the rate obtained from dividing the correct predictions by the total predictions generated from the model.

Bias – is the systematic error against certain groups due to or during the AI/ML decision-making process. To apply this definition in relation to the dataset, bias means the COMPAS recidivism classifier treats different races differently and misclassifies the entities more frequently in certain races over others. Specifically, when the Equal Accuracy option is selected, African Americans have a much higher False Positive rate than other races.

Fairness – is the ability to avoid discrimination during AI/ML decision-making. To apply this definition in relation to the dataset, Fairness means the COMPAS recidivism classifier treats different races equally, and the misclassification rate is the same across different races. Assuming that the false positive rates of African-American, Caucasian, Hispanic, Other, Asian and Native American are all the same, we can say that the COMPAS recidivism classifier is fair to those races through the lens of the false positive rate.

The differences between bias and Fairness in the AI/ML context can be summarized below: (1) Bias refers to the accuracy of the algorithm's predictions, while Fairness refers to the algorithm's ability to treat all groups equitably; (2) Bias is usually from the data used to train the algorithm, while Fairness is usually evaluated by the results generated from the algorithms; (3) An algorithm can be unbiased and unfair in the meantime or biased and fair at the same time.

2 DISCUSSION ON THE THREE RATES

False Positive should be most highly considered when wanting to mitigate bias. As defined in the previous section, ***False Positive*** refers to an entity being negative but misclassified as positive by the model. When a model makes the same number of predictions, a higher false positive rate means that more individuals who do not satisfy the conditions are marked as positive by the model. For the given dataset, the false positive rate for African-American is much higher than for other races, which means that African Americans are more likely to be predicted to be recidivists within two years by the model. The bias against African-Americans exists, and we want to mitigate such discrimination in the model. We want fewer African Americans to be misclassified as recidivists, i.e., a lower False Positive Rate.

False Positive should be most highly considered to ensure Fairness. According to the previous section, *F*airness stands for the model's ability to avoid discrimination during the AI/ML decision-making process. Ideally, the prediction results of the model should balance the False Positive rate and the False Negative rate so that the False Positive/False Negative rates of different races are the same to ensure Fairness. However, this ideal situation is challenging to achieve in reality, and the focus of fairness assurance should be on mitigating the African-American bias for the current dataset data characteristics. Therefore, False Positive should still be prioritized for the current dataset to ensure Fairness.

Yes, the rates I selected for bias mitigation and Fairness ensuring are the **same – False Positive**. As mentioned above, to generate relatively fair predictions, we should first mitigate the bias in the model. Even though in an ideal case, all three metrics – False Positives, False Negatives and Accuracy – should be considered, for the current data characteristics, False Positive is the one that should be given the most attention because there is a more significant variation in False Positive values across different races than in False Negatives that needs to be mitigated.

3 THRESHOLD VALUE CHANGE AND ITS IMPACT

Since the instructions select the False Positive value for Steps 3.3 and 3.4, this section will answer the questions for Steps 6 and 7 from the instructions. As shown in *Figure 1*, the thresholds are changed to mitigate the difference in False Positive values among each race to make them about the same. However, after

the threshold change, most of Race's False Negatives values remained the same or increased, especially for African-American (16.3% to 28.1%) and Caucasian (23.9% to 24.1%), while Accuracy values decreased overall - African-American from 65.1% to 60.8%, Hispanic from 68.2% to 67.8%, and other from 68.2% to 66.5%. This means that the change in the threshold has a negative impact on those races. Although Caucasian Accuracy increased by 0.1%, the difference was not significant. Because the sample size for Native American is too small (only 34 in count), the False Negatives and Accuracy changes are excluded from this discussion.

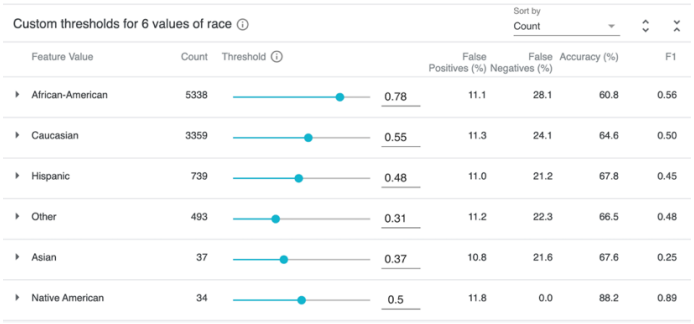

Figure 1—Change the threshold values to help mitigate bias based on the False Positives.

4 WHY MITIGATING BIAS AND ENSURING FAIRNESS AT THE SAME TIME IS A DIFFICULT TASK

Mitigating bias and ensuring Fairness at the same time is a difficult task. As discussed above, to ensure Fairness, ideally, we want to make False Positive, False Negative and Accuracy about the same. However, the reality is that when we mitigate a bias against a particular race, it will usually inevitably negatively impact other races. The discussion in the previous section is an example that can support this view.

5 DIFFERENT DATASET

If a different dataset is selected, my general definitions for the given terms would be similar, but their related meanings and my assessment could differ. The above definition relations and assessments with the dataset are mainly based on the data characteristics of the dataset. When the dataset changes, the corresponding data characteristics are supposed to change. Therefore, the corresponding definition relations and assessments should be different.