David Strube - dstrube3@gatech.edu
CS6603 - AI, Ethics, & Society
Written Critique: What-If Tool

1. **In your own words, define false negative, false positive, and accuracy rates.**

False negative: Incorrectly concluding that the thing you're looking for is not there.
False positive: Incorrectly concluding that the thing you're looking for is there.
Accuracy rate is the rate of accuracy. That is, the number of correct predictions divided by the number of total predictions.

2. **In your own words, define bias. Explain how bias can be applied to this dataset. Define fairness. Explain how fairness can be applied to this dataset. What is the difference between these two terms?**

Bias is the tendency to consistently learn the same wrong thing.

In this dataset, there is a bias for Caucasians (and other non-African American races) to be false negatives (i.e. determined **not** to be a recidivism risk which then turns out to be wrong); and there is a bias for African Americans to be false positives.

Fairness means that everyone has equal opportunities in all areas of life, regardless of age, sex, race, or any other characteristic that has no bearing on what is being applied for.

Given that there is racial bias clearly shown in the dataset, steps should be taken to make the recidivism system more fair.

In a way, bias and fairness are exact opposites: bias is bad, fairness is good. Bias may be difficult to avoid; fairness may be difficult to achieve. Bias should be minimized, fairness should be maximized.

3. **In the case of this dataset, which of the three rates (false negative, false positive, and accuracy) should be most highly considered when wanting to mitigate bias? Why?**

When wanting to mitigate bias, false positives should be most highly considered in the case of this dataset because false positives are equivalent to people being denied parole and kept in jail longer than they should be, which is worse than a false negative of people being granted parole when they shouldn't be.

4. **Of the three rates (false negative, false positive, and accuracy), which rate should be most highly considered to ensure fairness? Why?**

To ensure fairness, accuracy should be most highly considered because it is the metric that includes both false positives and false negatives in the denominator of its calculation.

5. **Are the rates you selected for bias and fairness the same? Why or why not?**

No, the rates changed because when optimizing for equal accuracy, a threshold per slice is optimized based on the specified cost ratio, ensuring the different slices achieve equal accuracy.

6. **In the What-If threshold window, change the threshold values to help mitigate bias based on the metric you chose in step 3.3 above. What happens to the rates for the other two terms? Do the corresponding results impact any groups negatively? (include a snapshot of the thresholds selected)**

Given that the metric I chose for step 3 above is false positives, I set the cost ratio (FP/FN) to 2. This reduced the false positives for all races, but also increased the false negatives for all races except for Native American which went down from 32.4 to 23.5. (Native American false positives were 0 both before and after the changes.) See appendix for before and after snapshots.

7. **In the What-If threshold window, change the threshold values to ensure fairness based on the metric you chose in step 3.4 above. What happens to them for the other two terms? Do the corresponding results impact any groups negatively? (include a snapshot of the thresholds selected) [Note: If you selected the same rates for bias and fairness, there is no need to rerun the analysis, just mention that here]**

Given that the metric I chose for step 4 is accuracy, I reverted the optimization strategy to equal accuracy. This reverted the false positives and false negatives to their previous values. This is a rerun of the previous analysis.

8. **Based on your assessment and definitions, is it a difficult task to mitigate bias and ensure fairness simultaneously? Why or why not?**

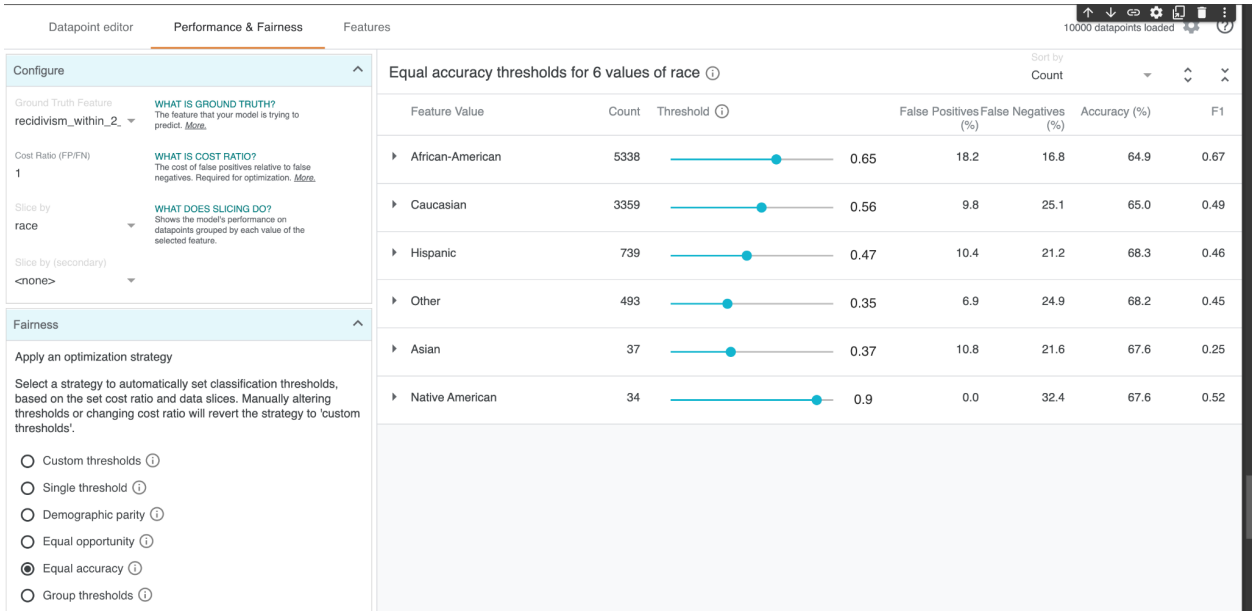
Yes, it is a difficult task to mitigate bias and ensure fairness simultaneously because it is a process that requires delicate balancing of several factors, and each tweak has multiple effects resulting in changes that affect multiple demographics. This may be best done with a system that graphs the results so as to visually represent the changes to make it easier to find the optimal settings.

9. **Would your assessment and definitions apply if a different dataset was selected? Why or why not?**

Hmm, that is a very interesting question. My guess is that the definitions would still apply because fairness, bias, false positives, false negatives, and accuracy are what they are regardless of the dataset; but the assessment might vary depending on the nature and complexity of the data.

Appendix

Step 6 before:



Step 6 after: