AI, Ethics, and Society

Written Critique Assignment #2

Readings:

- Chapter 10: Weapons of Math Destruction (The Targeted Citizen: Civic Life)
- "Machine Bias," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- "The What-If Tool: Code-Free Probing of Machine Learning Models," https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

Step 1: Navigate to Google's What-If Tool - https://pair-code.github.io/what-if-tool/

Step 2: Run the COMPAS Recidivism Classifier Notebook DEMO linked from the Google What-If Tool page -> https://colab.research.google.com/github/pair-code/what-if-tool/blob/master/WIT_COMPAS.ipynb

Under the "Invoke What-If Tool for test data and the trained models," you can see the same results that ProPublica found in their analysis by:

1. Selecting the "Performance & Fairness" tab
2. In "Ground Truth Feature" dropdown menu, select "recidivism_within_2_years"
3. In "Slice by" dropdown menu, select "race"
4. Under the "Fairness" option, select the "Equal accuracy" option

As you can see in the "Equal accuracy thresholds for 6 values of race" window, the different slices have very similar accuracy rates, but different false positive and false negative rates.

Step 3: Turn in a written critique (*no more than* 3 pages in length) in JDF Format addressing the following points:

1. In your own words, provide a definition for the terms false negative, false positive, and accuracy rates.
2. In your own words, provide a definition for bias. Explain how bias can be applied in relation to this dataset. Provide a definition for fairness. Explain how fairness can be applied in relation to this dataset. What is the difference between these two terms?
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?
4. Of the three rates (false negative, false positive, and accuracy), which rate should be most highly considered to ensure fairness? Why?
5. Are the rates you selected for bias and fairness the same? Why or why not?
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)
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 the rates 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*]
8. Based on your assessment and definitions, does it seem as if mitigating bias and ensuring fairness at the same time is a difficult task? Why or why not?

9. Do you think your assessment and definitions would apply if a different dataset was selected? Why or why not?

Step 4: Turn in a report (in PDF format) documenting your outputs for each question. The report should follow the JDF format. Reports that are not neat and well organized will receive up to a 10-point deduction. The file name for submission is GTuserName_WrittenCritique_2, for example, Joyner03_WrittenCritique_2.