Written Critique: What-If-Tool

Marcus J. Anderson

Online Masters of Computer Science

CS 6603: AI Ethics Society

Dr. Mahender Mandala

July 25, 2022

False negatives occur when the tested data is indicating a failed result, when in reality, it's actually a passing result. False positives are the same thing, expect the tested data indicates a passed result, when it's actually a failed result. We saw this a lot in the last couple years when COVID testing was at an all-time high. There were a variety of different methods to test if COVID was still in someone's system, but some methods were more accurate than others. Accuracy is how precise the data is in terms of correctness. For example, if I were to shoot an arrow at a target and hit the bullseye, that would be the most accurate result in this scenario.

By definition, bias is a prejudgment that's typically in favor of a privileged group such as a certain: age, race, religion, etc. In terms of this dataset, there may be more of a bias for a privileged group like the **African American** race, than the other races that exists within the data set. Fairness is the equal treatment between a privileged and unprivileged group, which would mean the racial groups that exists in the dataset would have equal treatment. The main difference between these two terms is that bias will highlight the favoritism of the privileged group over the unprivileged one, while fairness eliminates this favoritism and tries to make both groups equal.

I believe that the false positive rates would be the best possible way to help mitigate bias within the dataset. Typically, the privileged group has a higher favorable, or positive, count, which means that most of the data within this group is higher than, or equal to, the given threshold, as compared to the unprivileged group. If we start taking false positives into account, the unprivileged group, which tends to have a higher unfavorable count, would be then counted as favorable data, or falsely positive data. This would help close the bias gap between the privileged and unprivileged groups, thus mitigating the bias.

I believe the best way to ensure fairness between all slices within the dataset would be to use the accuracy rate. Since we're talking about fairness, bias should not be taken into account in this scenario. In order to be truly fair, we only care about the correctness of the data within the dataset, and the only rate that measures this is accuracy. Correctness equals facts, and the more factual a dataset is, the better chance we are able to ensure fairness between the slices of data.

I selected different rates for bias and fairness because even though a dataset can be formed to achieve both of these things, their core definitions still differ from one another. When we talk about bias in a dataset, this means that the privileged group that exists within it has preferential treatment over the unprivileged group. In order to mitigate this, I figured the best rate to use would be the false positive one, as it may help balance the bias against the unprivileged group as it holds more unfavorable (negative) data than privileged groups. For fairness in a dataset, this means that the data assigned to each slice is as correct as possible, true positives and true negatives. In order to achieve optimal fairness, I believe the accuracy rate is key, as it shows no favoritism between the privileged and unprivileged groups, just factual evidence. When facts are the main indicator, I believe that's the best way to display fairness within a dataset.

I was able to mitigate the bias to a 1.0 rating outcome using the **Disparate Impact** method, with the African-American slice as the privileged group, and the other five slices as the unprivileged group. In order to achieve this result, the threshold values in the unprivileged group needed to be lowered over 20%, while the threshold value of the privileged group remained unchanged. This also had an impact on both the accuracy and false negative rates, as both rates decreased a considerable amount in some slices. Since the privileged group's threshold value remained unchanged, I would say that no groups were negatively impacted as a result of mitigating the bias. Selected thresholds are shown at the end of the report.

I was able to achieve the top accuracy rate of each slice by actually only changing the threshold values of two slices, **Hispanic** and **Native American**. The other two terms changed a bit for these slices. For the Hispanic slice, the false positive rate decreased and false negative rate slightly increased, while only the false negative rate decreased in the Native American slice. Since there was a threshold value increased

within the Hispanic slice, I would say they got impacted the most by ensuring fairness throughout the dataset. Selected thresholds are shown at the end of the report.

From this assessment, I believe that mitigating bias and ensuring fairness throughout a dataset is possible, but very difficult. While I was able to mitigate the bias to a 1.0 outcome within the dataset, the accuracy rates were heavily impacted in a negative way. Whereas, when I set the threshold values to achieve the highest accuracy possible for each slice, the bias rate came in at .50, which is extremely lower than the rate the Disparate Impact method recommends.

Based on the information I've attained throughout this report, I believe that even if a different dataset was used, the same assessment and definitions I've mentioned would apply. Mitigating bias and ensuring fairness within a dataset mean the same thing to me, no matter what dataset is being analyzed. I do believe that there are some datasets out there that could achieve both bias mitigation as well as fairness, but the core definitions, as well as the considered rates, of both would always be the same.

Mitigated Bias - Threshold Screenshot:

Fairness - Threshold Screenshot: