Humana Mays Case Competition - Fairness in AI Guide

Overview

An important factor in gaining and keeping consumer trust is to ensure that the algorithms using their data operate fairly and without bias. All data which involves the society we live in, including healthcare data, is encoded with societal biases. Without a careful eye, these biases can be mirrored or even amplified in the algorithms which utilize the data. Specifically, these biases risk causing harm to individuals who have historically already been harmed by these same societal biases. While there are no regulatory requirements today, Humana has been investing in technology and resources to ensure that we can take a close look at our algorithms to ensure our members and associates are treated fairly.

Even though a model doesn't have a sensitive feature such as race as an input doesn't make it immune to having racial bias. Biases are often encoded within the data in unexpected ways, and data involving society will mirror the biases of the society. Data science models have a tendency to amplify such bias. Take zip code as an example: even though race data isn't directly part of a zip code, it has a strong correlation since many neighborhoods are still essentially segregated. While this is a strong example, similar correlations are present throughout societal data - especially healthcare data.

It is also important to consider that fairness is a complicated subject even with software to assist in the analysis. There are more than 30 well regarded metrics available for considering whether an algorithm is fair. As a result, Humana has been developing best practice guidelines for our data scientists, engineers, and subject matter experts to help navigate fairness in the most scenarios.

Equal Opportunity

For this challenge, we will be measuring fairness using Equal Opportunity metric: group should get the positive outcome at equal rates, assuming that people in this group qualify for it. In other words, the **predicted positive rate** should be proportional to that of the ground truth positive rate.

Disparity Ratio

For each group within a sensitive variable (***RACE***, ***SEX***), disparity ratio DR is defined as:

$$DR = \frac{S_n}{S_0}$$

Where S_n is the scoring metric (AUC, positive rate, etc) for each class and S_0 is the scoring metric for the reference group. The reference group is defined as the privileged group (i.e. for Medicare data Race: White, Sex: Male) within a sensitive variable class. For this competition, the scoring metric will be ***predicted positive rate***. Note: This is not the true positive rate, but rather the sum of true positives and false positives.

In order to get a single score, the total disparity ratio should be calculated as the average (with total number of classes in all sensitive variables *N*) across all sensitive variables *SV* the sum of each disparity ratio capped at 100%:

$$Disparity\ Score = \frac{\sum_{SV} \sum_{S} \min\left(\frac{S_n}{S_0}, 1\right)}{N}$$

Disparity Score Weighting

The disparity score will determine a weight for the overall model performance as described below:

Disparity Score	***Weight***
90 - 100	1.0
80 - 89	0.95
70 - 79	0.90
>70	0.85