Responsible Machine Learning
Assignment 3
10 points

June 8, 2022

In Assignment 3, you will work with your group to test and remediate interpretable machine learning (ML) models for algorithmic discrimination following the instructions below. A template has been provided as an example of testing and remediating interpretable models for algorthmic discrimination. For those of you who use Python virtual environments, a basic `requirements.txt` file is also available for the template.

Please let me know immediately if you find typos or mistakes in this assignment or related materials.

1 Test Models for Discrimination using AIR.

Test your best performing interpretable ML model across major demographic groups for discriminatory outcomes using adverse impact ratio (AIR), to include Asian, Black, and White people, and males and females. Cells 13–17 of the template provide an example for this analysis.

2 Remediate Discovered Discrimination.

You will likely discover problematic discrimination against Black people. (Note that this is real data and that these are real people paying more for the American Dream of home ownership.) Your notebook must propose at least two discrimination remediation techniques. The template presents a simple approach of changing probability cutoffs in cells 18–19. The template also presents a more in-depth remediation approach in cells 20–24, wherein a large grid search occurs across many random hyperparameter and feature sets in hopes of finding a less discriminatory model that retains high quality. (This grid search requires considerable time to complete.) You may try other remediation processes that you find online, just remember that they may not be legally viable in the regulated U.S. fair lending and employment contexts. However you remediate your model, you must also show that you have not reduced any other group's AIR score below 0.8.

3 Submit Code Results.

Your deliverable for this assignment is to update your group's GitHub repository to reflect this discrimination testing and remediation analysis. The first round of AIR test results are worth 3 points and correct application of remediation approaches is worth another 3 points (which will involve more AIR tests). The remaining 4 points for the assignment will be granted based on the AUC of your remediated model. The higher the AUC for your final model, with all AIR tests above 0.8 and retaining a selective 0.17 cutoff, the higher your grade will be. As a benchmark, the template approach achieves 0.7847 AUC with a minimum AIR of 0.8043 at a 0.17 cutoff.

Your deliverables are due Wednesday, June 11th, at 11:00 AM ET.

Note that you may also improve Assignment 1 scores throughout the Summer I Session to improve your ranking, your Assignment 1 grade, and your final project grade.