Responsible Machine Learning
Assignment 3
10 points

November 6, 2026

In Assignment 3, work alone or with a group to test and remediate explainable machine learning (ML) models for systemic bias. You may use AI tools. Submit your code and predictions to the class evaluation server at a TBD URL.

1 Test Models for Bias using AIR.

Test your best-performing explainable ML model across major demographic groups for discriminatory outcomes using adverse impact ratio (AIR), including Asian, Black, Hispanic, and White people, and males and females. Report the AIR for each group against an explicitly identified reference group.

2 Remediate Discovered Discrimination.

Propose and evaluate at least two techniques for remediating any disparities you find. For each approach, describe its effect on model performance and group AIR. You may consider changing probability cutoffs, changing model features or parameters, or applying another remediation method. Explain the limitations of the methods you use, including their suitability for regulated U.S. fair lending and employment contexts. Check that remediation does not reduce any other group's AIR below 0.8.

3 Submit Code Results.

Submit your analysis, code, and predictions in a public GitHub repository, or a private repository shared with the instructor (jphall663). Include initial AIR results (3 points), at least two remediation approaches with follow-up AIR results (3 points), and final model results (4 points). Models whose reported AIR values are at least 0.8 for all groups will be eligible for the class leaderboard. The final 4 points are based on your model's rank among eligible submissions. Aim to improve your rank while preserving the AIR threshold, explain the decision cutoff used, and discuss fairness trade-offs.
Your deliverables are due: TBD.