Responsible Machine Learning

Assignment 4

10 points

November 6, 2026

In Assignment 4, work alone or with a group to red-team the best model you've developed so far. Treat that model as a black box when conducting the attacks. You may use AI tools, but you are responsible for validating and understanding your results.

1 Conduct a white-hat model extraction attack.

Use queries to the Assignment 3 model to train and document a surrogate that approximates its predictions. Explain your query strategy, the surrogate model, and how well it reproduces the target model's behavior.

The model extraction attack is worth 3 points.

2 Find adversarial examples for your model.

Use the Assignment 3 model and your analysis to create at least one valid input that produces a low prediction and at least one that produces a high prediction. Describe the changes made to each input and the resulting predictions.

The adversarial examples are worth 3 points.

3 Submit Code and Results.

Submit your code and predictions to the class evaluation server at a TBD URL. Keep your work in a public GitHub repository, or a private repository shared with the instructor (`jphall663`). Assignment 3 models whose reported AIR values are at least 0.8 for all protected groups will be eligible for the class leaderboard. The model ranking is worth 4 points. Aim to improve your model's leaderboard position while preserving the AIR threshold, and report your rank and AIR values for all protected groups alongside your attack results.

In the real world, after performing this red-teaming exercise, you would want to contact your manager and your organization's IT security team to discuss any discovered vulnerabilities. Countermeasures to discuss with business and IT colleagues may relate to authentication on the vulnerable model API endpoint, throttling/rate-limiting of the vulnerable model API and monitoring the model's production scoring queue for random data and training data, if possible.

Your deliverables are due: TBD.