HW1 xAI Drochomirecki

Wojciech Drochomirecki

October 2024

1 Task (2)

For the first model, I chose to train a random forest on the following set of features: ”sex,” ”race,” ”juv_fel_count,” ”juv_misd_count,” ”juv_other_count,” ”priors_count,” ”c_charge_degree,” and ”decile_score.” I encoded categorical features as one-hot vectors and did not provide any protected features for prediction to reduce bias. The trained model achieved 62% accuracy. The table below shows the calculated fairness metrics:

Statistical Parity	**Equal Opportunity**	**Neg Predictive Parity**	**Pos Predictive Parity**
178	138	123	158

Table 1: Fairness coefficients for the first model

Almost all statistics fail to meet the four-fifths rule, which suggests significant bias in the model.

2 Task (3)

I compared the performance of the random forest with different architectures trained on the same data. For comparison, I chose XGBoost and logistic regression. Both models achieved comparable results, with XGBoost at 65% accuracy and logistic regression at 66%. The table below presents a comparison of fairness metrics for these models with the previously trained random forest:

Statistical Parity	**Equal Opportunity**	**Neg Predictive Parity**	**Pos Predictive Parity**
178	138	123	158
177	130	115	161
184	124	111	162

Table 2: Fairness coefficients comparison

The results show that the choice of model type does not significantly impact fairness. All models have coefficients far from the ideal proportion.

3 Task (4)

I decided to use data balancing as a mitigation technique to improve the fairness of the trained models. I upsampled positive samples in the underrepresented class to balance the classes, then retrained all three models on the extended dataset. The results are shown in the table below:

Statistical Parity	Equal Opportunity	Neg Predictive Parity	Pos Predictive Parity
156	131	132	156
159	135	134	147
157	120	121	162

Table 3: Fairness coefficients after mitigation

While the results are still not ideal, there is a significant improvement, especially in metrics that were previously much higher. After mitigation, most of the metrics are either compliant with or close to meeting the four-fifths rule.

4 Task (5)

The table below shows the accuracy of the models before and after mitigation:

Model	Before Mitigation	After Mitigation
Random Forest	0.625	0.617
XGBoost	0.654	0.647
Logistic Regression	0.666	0.665

Table 4: Model accuracy before and after mitigation

This data shows that increasing fairness comes at the cost of accuracy. While the drop is noticeable, one can argue that in reality, this helps with model generalization under the assumption that the correlation between the response and the sensitive feature is a result of dataset bias.