HW3: Detecting and Mitigating Algorithmic Bias – Part 3

Bias Mitigation

I evaluated two fairness approaches using sex as the sensitive attribute. The first method resampled the training data so each sex/income group was represented equally. The second used a fairness-constrained model that enforced demographic parity during training.

Results

Table 1: Performance and Fairness Before and After Mitigation

Metric	**Baseline**	**Resampling**	**Fairness-Constrained**
Accuracy	0.846	0.808	0.827
Precision	0.737	0.581	0.736
Recall	0.589	0.804	0.472
F1	0.654	0.675	0.575
DP Diff	0.176	0.196	**0.007**
EO Diff	0.114	**0.041**	0.282

Evaluation and Trade-Offs

The baseline model performed well overall, but its fairness metrics showed clear disparities between men and women (DP diff = 0.176, EO diff = 0.114).

Resampling improved equal opportunity the most (0.114 → 0.041) and boosted recall. However, accuracy and precision dropped, and demographic parity did not improve. This method makes the model better at treating qualified individuals from both groups equally, but it produces more false positives.

The **fairness-constrained model** nearly eliminated demographic parity difference (0.176 → 0.007) while keeping precision high. The trade off is a large drop in recall and worse equal opportunity (0.282). This shows that enforcing equal prediction rates can make the model less consistent at identifying true positives.

Overall, both methods reduce bias but in different ways. Improving one fairness measure usually harms another, and both approaches cost some performance. The best choice depends on which fairness goal, equal outcomes or equal true positive rates is more important for the application.