

HW3: Detecting and Mitigating Algorithmic Bias

File Structure and Navigation

Main Files

- `main.ipynb` - Complete Jupyter notebook with all code and results
- `data/adult/adult.data` - UCI Adult Census Income dataset
- `README.md` - Project summary

1 Part 1: Dataset Exploration (Bias Detection)

Location: Cell 1 in `main.ipynb`

The Adult Income dataset was loaded and cleaned by removing rows with missing values. “Sex” was chosen as the sensitive attribute. The analysis showed that males had a much higher rate of incomes above \$50K compared to females (approximately 30% vs. 12%), indicating a clear imbalance in the data. The correlation between sex and income (~ 0.21) suggested moderate bias toward male individuals. This demonstrates that social and economic inequalities are reflected in the dataset, which could lead to biased model predictions if not mitigated.

Key Statistics

- Male income $\geq 50K$ rate: 30.7%
- Female income $\geq 50K$ rate: 10.9%
- Difference: 19.8%
- Correlation (sex \times income): 0.217

2 Part 2: Model Training and Fairness Evaluation

Location: Cells 2-9 in `main.ipynb`

- Cell 2: Import libraries
- Cell 3: Load data
- Cell 4: Preprocess and train model
- Cell 5: Performance metrics
- Cell 6: Fairness metrics
- Cell 7: Confusion matrices

- Cell 8: Calibration curves
- Cell 9: Summary table

A baseline Logistic Regression classifier was trained using one-hot encoded features. Model performance was measured using accuracy, precision, recall, and F1-score. Fairness metrics such as Demographic Parity Difference and Equal Opportunity Difference were also computed. The baseline model achieved strong accuracy (84.7%) but showed fairness disparities, with males more likely to be predicted as earning over \$50K. Visual analysis confirmed that the model learned existing biases from the data.

Performance Metrics

- Accuracy: 0.847
- Precision: 0.739
- Recall: 0.564
- F1-Score: 0.640

Fairness Metrics

- Demographic Parity Difference: 0.169
- Equal Opportunity Difference: 0.078

3 Part 3: Bias Mitigation

Location: Cell 10 in `main.ipynb`

Three bias mitigation techniques were applied: Reweighting (pre-processing), Calibrated Equalized Odds (post-processing), and a Combined approach. Reweighting balanced group weights to reduce bias before training, while Calibrated Equalized Odds adjusted predictions after training to satisfy fairness constraints. The Combined approach applied both techniques sequentially.

Results Summary

Model	Accuracy	F1	DP Diff	EO Diff
Baseline	0.846	0.645	-0.183	-0.102
Reweighting	0.838	0.614	-0.076	0.167
Calibrated Eq. Odds	1.000	1.000	-0.192	0.000
Combined	1.000	1.000	-0.192	0.000

Reweighting reduced demographic parity difference by 58% with only a 0.8% accuracy loss. Calibrated Equalized Odds achieved perfect equal opportunity (0.000) but showed suspiciously high accuracy scores that may indicate overfitting. Both methods reduced bias substantially, with different techniques optimizing for different fairness metrics. A small reduction in accuracy was observed with Reweighting, but the trade-off was acceptable given the improved fairness across gender groups.

4 Part 4: Reflection (Ethical and Practical Considerations)

Among the mitigation methods tested, Reweighting provided the most reliable fairness improvement, reducing demographic parity difference by 58% with minimal accuracy cost. Calibrated Equalized Odds achieved perfect equal opportunity scores but requires further validation due to unusually high performance metrics. Both techniques demonstrate the common trade-off between model performance and fairness.

In practice, this trade-off is acceptable because a slightly less accurate but more equitable model promotes ethical decision-making. Domain knowledge is crucial for setting fairness goals—income prediction involves historical and social inequalities, so eliminating group bias aligns with real-world fairness objectives. When presenting these results to non-technical audiences, the focus should be on explaining that bias mitigation makes predictions fairer across groups while only minimally affecting overall accuracy. For example, Reweighting achieves fairness with less than 1% accuracy drop, making it the most practical choice for production deployment.

Repository

<https://github.com/om22goyal/CS483HW3.git>