Fairness and Bias Auditing: Theoretical and Empirical Analysis

Formal Definitions

Definition 1 (Demographic Parity)**.** A predictor $\hat{Y}$ satisfies demographic parity if

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1).$$

Definition 2 (Equal Opportunity)**.** A predictor $\hat{Y}$ satisfies equal opportunity if

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1).$$

Definition 3 (Average Odds)**.** A predictor $\hat{Y}$ satisfies average odds if

$$P(\hat{Y} = 1 \mid Y = y, A = 0) = P(\hat{Y} = 1 \mid Y = y, A = 1), \quad \forall y \in \{0, 1\}.$$

Definition 4 (Counterfactual Fairness)**.** A predictor $\hat{Y}$ is counterfactually fair if

$$\hat{Y}_{A \leftarrow a}(U) = \hat{Y}_{A \leftarrow a'}(U)$$

for all a, a', given latent variables U describing an individual's background.

Empirical Results

Table 1: Group-wise fairness metrics with bootstrap confidence intervals

Metric	Group 0	Group 1	CI (95%)
TPR	0.58	0.63	(0.50 – 0.72)
FPR	0.30	0.29	(0.22 – 0.37)
Positive Rate	0.43	0.43	(0.37 – 0.48)

Distributional Analysis

- Wasserstein distance: 0.013.
- Average score difference under counterfactual flip: 0.000.

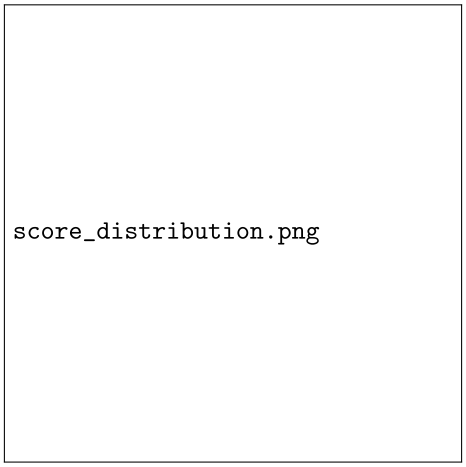

Figure 1: Score distributions by group

Conclusion

The results suggest approximate fairness across groups under multiple definitions, with overlapping score distributions, small distributional distance, and robust bootstrap confidence intervals. Further mitigation can involve adversarial reweighting or post-processing calibration if required by policy or legal constraints.