COMPAS Fairness Audit Report

1. Introduction

The COMPAS dataset is widely used to predict recidivism risk but has been criticized for racial bias. This report audits the dataset using IBM's AI Fairness 360 toolkit, evaluates disparities across groups, and applies remediation techniques to improve fairness.

2. Baseline Audit Results

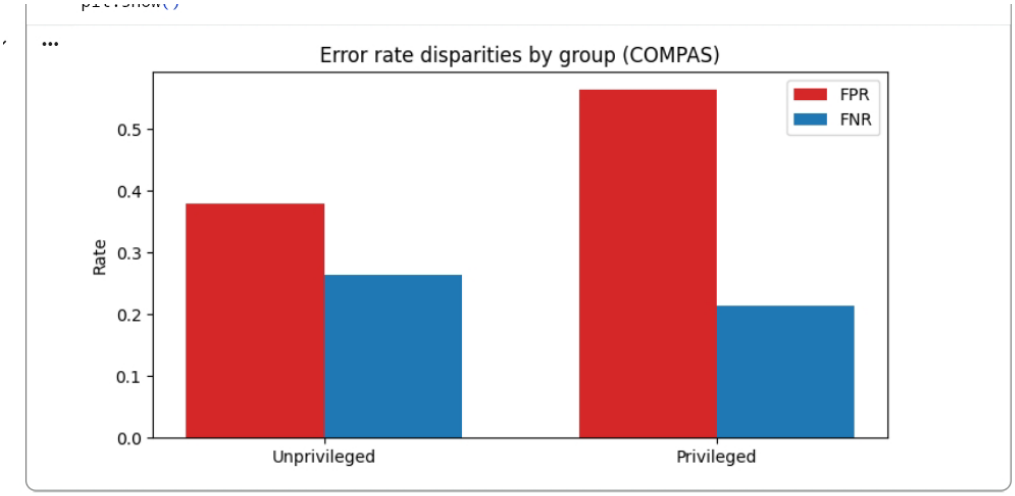
Initial evaluation with a logistic regression classifier revealed significant disparities between unprivileged (African-American) and privileged (Non-African-American) groups.

- Disparate Impact: Deviates from parity (ideal ≈ 1).

- Equal Opportunity Difference: Shows unequal true positive rates.

- Error Rates: Higher false positives for unprivileged groups, meaning they are more often incorrectly labeled "high risk."

Figure 1: Error Rate Disparities by Group (COMPAS)



This chart compares false positive rates (FPR) and false negative rates (FNR) between unprivileged and privileged groups. The disparities highlight that unprivileged groups experience higher false positives, demonstrating unequal treatment and potential harm in real-world

applications.

3. Remediation with Reweighing

We applied the Reweighing algorithm from AIF360, which adjusts instance weights to balance group influence during training.

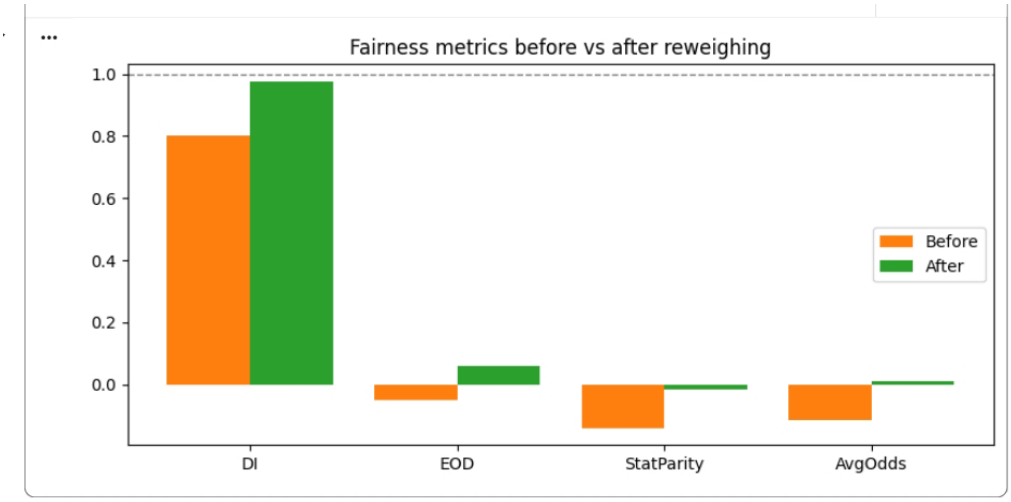
Post-remediation results showed improvements:

- Disparate Impact: Moved closer to parity.

- Equal Opportunity Difference: Narrowed.

- Average Odds Difference & Statistical Parity Difference: Reduced gaps.

Figure 2: Fairness Metrics Before vs After Reweighing



This chart shows key fairness metrics — Disparate Impact (DI), Equal Opportunity Difference (EOD), Statistical Parity Difference, and Average Odds Difference — before and after reweighing. The “After” bars move closer to parity, demonstrating that fairness-aware preprocessing can mitigate systemic disparities.

4. Discussion

Although reweighing improved fairness metrics, disparities remain. This highlights that fairness

is not a one-off fix but requires continuous monitoring, diverse data, and governance. Remediation should be multi-pronged:

- Fairness-aware training and threshold calibration.

- Independent audits and transparent documentation.

- Human-in-the-loop review for high-stakes decisions.

- Clear explanations and contestability for individuals affected.

5. Conclusion

The COMPAS audit demonstrates both the risks of biased AI and the potential of fairness-aware methods to reduce harm. Ethical AI requires ongoing evaluation, inclusive oversight, and accountability to align predictive systems with societal values.