Bias Audit Report

COMPAS Recidivism Dataset This audit investigates racial bias in the COMPAS recidivism risk scores using the AI Fairness 360 toolkit. We specifically analyzed disparities between the privileged group (Caucasian) and the unprivileged group (African-American) using the false positive rate (FPR) as the fairness metric. A high FPR means a person is incorrectly predicted to reoffend when they would not, which can lead to unjust detention or parole denial. Baseline Evaluation: Before applying any fairness mitigation, the FPR for Caucasians was 0.611, while for African-Americans it was significantly lower at 0.290. This resulted in a negative FPR gap of -0.321, indicating that Caucasians were far more likely to be incorrectly flagged as high risk compared to African-Americans. While this direction of bias may seem counterintuitive to widely reported COMPAS disparities, it reflects how the FPR is behaving on this filtered subset and classifier setup. It also underscores the importance of using data-driven metrics over assumptions. Bias Mitigation with Reweighing: To address this disparity, we applied the Reweighing algorithm, which adjusts instance weights during training to balance the treatment of different demographic groups. After mitigation, the FPR for Caucasians dropped to 0.377, and increased slightly for African-Americans to 0.423. The resulting FPR gap narrowed significantly to +0.046, indicating a substantial reduction in bias. Conclusion & Recommendations:

The results demonstrate that pre-processing techniques like Reweighing can reduce bias in recidivism predictions without entirely sacrificing predictive structure. However, practitioners should also evaluate fairness across other metrics like false negative rate and disparate impact. In production, fairness-aware models should be paired with transparent communication, regular audits, and community oversight, particularly in high-stakes settings like criminal justice.