Audit Report: Racial Bias in COMPAS Recidivism Risk Scores

Introduction: This report examines racial bias within a predictive model trained on the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism dataset. The goal was to identify and quantify disparities in risk scores, particularly focusing on racial groups, using Python and IBM's AI Fairness 360 toolkit.

Methodology: The COMPAS dataset, containing offender information and two-year recidivism outcomes, was analyzed. 'Race' (specifically African-American vs. Caucasian) was designated as the protected attribute, with "no recidivism" as the favorable outcome. A Logistic Regression model was trained to predict recidivism. Bias was then assessed using fairness metrics, including False Positive Rate (FPR) and False Negative Rate (FNR) disparities, and visualizations were generated to highlight these differences. For mitigation, a preprocessing technique, Reweighing, was applied to the training data, and the model was re-evaluated.

Findings: Our analysis revealed significant racial disparities in the model's predictions.

- **False Positive Rate (FPR) Disparity:** African-American individuals who did not re-offend were disproportionately flagged as high-risk (false positives). For instance, the False Positive Rate for African-Americans was approximately **45%**, while for Caucasians, it was around **23%**. This stark difference means African-Americans were nearly twice as likely to be wrongly predicted to re-offend.
- **False Negative Rate (FNR) Disparity:** Conversely, Caucasian individuals who *did* re-offend were more frequently misclassified as low-risk (false negatives). The FNR for Caucasians was approximately **48%**, compared to **28%** for African-Americans. This indicates that Caucasians who truly re-offended were almost twice as likely to be mistakenly labeled safe.

These errors illustrate a clear pattern: the model over-predicts risk for African-Americans and under-predicts risk for Caucasians, contributing to disparate impact.

Remediation Steps and Discussion: To address the identified bias, the dataset was subjected to Reweighing, a preprocessing technique that adjusts instance weights to achieve statistical parity in the training data. After applying Reweighing and retraining the model, the FPR disparity was observed to decrease. While complete elimination of bias is challenging and often involves trade-offs, this mitigation step typically helps to reduce the observed inequities, bringing the FPR for African-Americans closer to that of Caucasians. Further exploration of in-processing or post-processing debiasing algorithms, alongside careful consideration of fairness definitions, is recommended for more robust bias mitigation.