Bias Audit Report: Racial Fairness Analysis of the COMPAS Dataset

This report presents an audit of racial bias within the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset using the IBM AI Fairness 360 (AIF360) toolkit. The COMPAS dataset is frequently used in the U.S. criminal justice system to predict the likelihood that a defendant will reoffend. However, the use of such predictive tools has raised ethical concerns, particularly around racial discrimination and fairness.

Our analysis focused on measuring potential disparities between two racial groups: Caucasians (privileged group) and African-Americans (unprivileged group). To assess fairness, we utilized the BinaryLabelDatasetMetric class from AIF360, which provides key fairness metrics including:

- Statistical Parity Difference: -0.097
- Disparate Impact: 0.84

A statistical parity difference close to 0 and a disparate impact close to 1 are considered fair. Our findings show that African-American defendants are approximately 9.7% less likely to be classified into favorable outcome categories than their Caucasian counterparts. A disparate impact below 0.8 generally indicates a potential violation of fairness standards, meaning our result (0.84) is borderline and worth addressing.

We also visualized the base rates (the proportion of individuals labeled as low risk) across the two groups. The bar chart showed a clear imbalance, with the privileged group enjoying a higher base rate of favorable outcomes. This reinforces the conclusion that racial bias exists in the dataset prior to modeling.

Remediation Strategy

To mitigate this bias, we applied the Reweighing technique, a pre-processing algorithm that assigns different weights to data points based on their group membership. This method helps balance the representation of privileged and unprivileged groups in training data, without altering feature values or labels directly. After reweighing, the transformed dataset reflects a fairer distribution of favorable outcomes, laying the groundwork for more equitable machine learning models.

Conclusion

Our audit confirms measurable racial disparities in the COMPAS dataset. Applying fairness-aware pre-processing methods like Reweighing is a crucial step toward reducing bias in automated decision systems. However, technical remediation alone is not sufficient—

organizations must also consider the social and ethical implications when deploying predictive tools in sensitive domains such as criminal justice.

Post-Mitigation Analysis

The following chart compares the base rates of favorable outcomes (e.g., low-risk predictions) between privileged and unprivileged racial groups before and after applying the Reweighing technique.

Before mitigation, there was a notable disparity: Caucasian individuals had a significantly higher base rate of favorable outcomes compared to African-American individuals. After applying Reweighing, the base rates became more balanced, reducing the bias in outcome distribution.

This visualization confirms the effectiveness of Reweighing as a fairness-aware pre-processing step. While the rates are not perfectly equal, the gap between groups has narrowed, indicating improved fairness in the dataset.