COMPAS Bias Audit Report

Title: Racial Bias Audit of COMPAS Risk Scores Using AI Fairness 360

Introduction

This report presents an audit of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset using IBM's AI Fairness 360 toolkit. The primary objective was to detect and quantify racial bias in predicted risk scores, particularly between African-American and Caucasian individuals.

The COMPAS dataset contains risk scores used to predict the likelihood of recidivism within two years. These scores have faced scrutiny for potentially introducing racial bias in criminal justice decision-making.

Methodology

- The dataset (`compas-scores-two-years.csv`) was preprocessed to retain key features: `race`, `age`, `priors_count`, `decile_score`, and `two_year_recid`.
- Using the `decile_score` as a proxy classifier, we simulated predictions by labeling individuals with scores ≥5 as "high risk" (predicted = 1).
- We defined **privileged group** as `Caucasian` and **unprivileged group** as `African-American`.
- The IBM AI Fairness 360 (AIF360) toolkit was used to construct `BinaryLabelDataset` objects for both true and predicted labels.
- Fairness metrics were calculated, and a visualization of predicted outcomes by race was generated.

Fairness Metrics Summary

Metric	**Value**	**Interpretation**
Statistical Parity Difference	0.2402	24% more likely for African-Americans to be labeled high-risk
Disparate Impact	1.6902	Higher rate of adverse prediction for African-Americans
Equal Opportunity Difference	0.1974	Higher true positive rate for African-Americans
Average Odds Difference	0.2056	Notable imbalance in true/false positives
False Positive Rate Difference	0.2139	African-Americans more likely to be wrongly labeled high-risk

These results suggest systemic bias in COMPAS scores, with African-Americans facing a significantly higher likelihood of being incorrectly classified as high risk.

Visualization

A bar chart of predicted outcomes showed a visibly higher number of African-Americans labeled as high-risk compared to Caucasians, reinforcing the numeric results.

Recommendations

To address this bias, the following steps are recommended:

- Apply fairness-aware techniques (e.g., **reweighing**, **adversarial debiasing**) during model training.
- Use **post-processing algorithms** like **calibrated equalized odds** to adjust predictions fairly.

- Ensure ongoing **bias audits** in real-world predictive systems, especially in high-stakes sectors like justice and healthcare.

Conclusion

This audit confirms the presence of racial bias in the COMPAS dataset's predictive outcomes. Leveraging fairness toolkits like AIF360 can help make these biases visible, measurable, and ultimately correctable. Ethical use of AI requires transparency, fairness, and continuous evaluation.