COMPAS Recidivism Risk Assessment: Racial Bias Audit Report

Executive Summary

This audit analyzes the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset for racial bias in recidivism risk scoring. Using IBM's AI Fairness 360 toolkit, we examined disparities between African-American and Caucasian defendants across multiple fairness metrics.

Key Findings

1. Disparate Impact Detected

The analysis reveals significant racial bias in COMPAS risk assessments. The disparate impact ratio falls outside the acceptable 0.8-1.2 range, indicating that African-American defendants receive systematically different risk classifications compared to Caucasian defendants.

2. False Positive Rate Disparity

African-American defendants experience substantially higher false positive rates, meaning they are more frequently misclassified as high-risk when they do not recidivate. This disparity leads to harsher sentencing and denial of bail for individuals who pose no actual risk.

3. Statistical Parity Violation

The statistical parity difference exceeds acceptable thresholds, demonstrating that positive outcome rates (being classified as low-risk) differ significantly between racial groups, even when controlling for actual recidivism.

Remediation Recommendations

Immediate Actions:

- **Implement Reweighing**: Apply preprocessing bias mitigation using reweighing algorithms to adjust training data weights, reducing group-based disparities while maintaining predictive accuracy.

- **Recalibrate Risk Thresholds**: Establish race-specific decision thresholds or use threshold optimization techniques to equalize false positive and false negative rates across groups.

Systemic Changes:

- **Remove Proxy Variables**: Audit features for proxies of race (zip code, neighborhood characteristics) and eliminate or transform problematic predictors.

- **Adversarial Debiasing**: Train models using adversarial learning to simultaneously optimize for accuracy and fairness metrics.

- **Human-in-the-Loop Review**: Mandate manual review for high-risk classifications, particularly for minority defendants, ensuring algorithmic decisions undergo judicial scrutiny.

Monitoring Framework:

Establish continuous fairness monitoring with quarterly audits tracking disparate impact, false positive rate parity, and equalized odds. Implement automated alerts when fairness metrics drift beyond acceptable bounds.

The COMPAS system requires immediate intervention to prevent perpetuation of systemic racial bias in criminal justice decisions.