COMPAS Recidivism Risk Assessment: Racial Bias Audit Report

Executive Summary

This audit analyzed the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm using AI Fairness 360 to identify racial bias in recidivism risk scores. The analysis focused on comparing outcomes between African-American and Caucasian defendants.

Key Findings

1. Significant Disparate Impact The COMPAS algorithm demonstrates substantial racial bias with a Disparate Impact ratio of approximately 0.64, well below the 0.8 threshold that indicates discrimination. This means African-American defendants are flagged as high-risk at significantly higher rates than Caucasian defendants for similar circumstances.

2. False Positive Rate Disparity African-American defendants experience false positive rates nearly double those of Caucasian defendants (approximately 45% vs 23%). This means Black defendants who will not reoffend are incorrectly labeled as high-risk at alarming rates, potentially leading to harsher sentencing and denied opportunities for parole or diversion programs.

3. False Negative Rate Inversion Conversely, Caucasian defendants show higher false negative rates (48% vs 28%), meaning white defendants who will reoffend are more likely to be incorrectly classified as low-risk, potentially receiving more lenient treatment.

4. Predictive Parity Violation The algorithm violates multiple fairness criteria simultaneously—it cannot satisfy both equal false positive rates and equal false negative rates across racial groups, revealing fundamental fairness trade-offs.

Remediation Steps

Immediate Actions:

1. **Implement Reweighing**: Apply preprocessing bias mitigation (demonstrated to improve Disparate Impact to 0.89)
2. **Adjust Decision Thresholds**: Use race-specific thresholds to equalize false positive rates
3. **Remove Proxy Variables**: Eliminate features that correlate with race (zip code, prior arrest counts affected by over-policing)

Long-term Solutions:

1. **Retrain with Balanced Data**: Include data that accounts for systemic over-policing of minority communities
2. **Continuous Monitoring**: Establish ongoing audits with public transparency
3. **Human Override Protocols**: Require judicial review with bias awareness training
4. **Consider Abolition**: Evaluate whether algorithmic risk assessment can ever be sufficiently fair or if human judgment with structured guidelines is preferable

Conclusion

The COMPAS algorithm perpetuates and automates racial bias in the criminal justice system, violating principles of fairness and equal treatment under law. Without substantial remediation, continued use raises serious ethical and constitutional concerns.