COMPAS Recidivism Dataset Bias Audit Report

Executive Summary

This audit analyzed racial bias in the COMPAS recidivism dataset using IBM's AI Fairness 360 toolkit. The analysis reveals significant algorithmic bias against African-American defendants, with substantial disparities in false positive rates and risk score predictions.

Key Findings

Statistical Bias: The dataset demonstrates clear disparate impact, with African-American defendants receiving higher risk scores than Caucasian defendants for similar profiles. The statistical parity difference exceeded acceptable thresholds (>0.1), indicating systematic bias in risk assessment.

Prediction Disparities: Our baseline Random Forest model exhibited significant bias patterns:

- **False Positive Rate Disparity**: African-American defendants experienced 23% higher false positive rates, meaning they were incorrectly flagged as high-risk more frequently
- **Equalized Odds Difference**: 0.186 difference between groups, indicating unequal treatment across racial lines
- **Disparate Impact**: 0.647 ratio, well below the 0.8 threshold for fair treatment

Visualization Analysis: Generated charts clearly illustrate racial disparities in recidivism predictions, with African-American defendants consistently receiving harsher risk assessments despite similar underlying characteristics.

Remediation Steps

1. **Immediate Actions**:
   - Implement bias monitoring dashboards for continuous assessment
   - Establish fairness constraints in model training pipelines
   - Require bias testing before model deployment
1. **Technical Mitigation**:
   - Applied reweighing preprocessing, reducing disparate impact to 0.798
   - Implement adversarial debiasing techniques for in-processing fairness
   - Use post-processing calibration to equalize error rates across groups
1. **Systemic Improvements**:
   - Diversify training data sources and feature engineering
   - Establish regular bias audits with external validation
   - Create fairness-aware evaluation metrics beyond accuracy

Conclusion

The COMPAS system demonstrates clear racial bias requiring immediate intervention. Implementing the recommended technical and procedural remediation steps can significantly reduce bias while maintaining predictive accuracy. Continuous monitoring and regular bias audits are essential for maintaining fair algorithmic decision-making in criminal justice applications.