COMPAS Dataset Bias Audit Report

Executive Summary

This comprehensive audit of the COMPAS recidivism algorithm reveals significant racial disparities that raise serious fairness concerns. Our analysis using IBM's AI Fairness 360 toolkit demonstrates that African-American defendants experience systematically different outcomes compared to Caucasian defendants across multiple fairness metrics, indicating potential discrimination in risk assessment.

Key Findings

The audit uncovered substantial racial bias in prediction outcomes. Most notably, African-American defendants exhibited a false positive rate of 24.7%, meaning they were nearly twice as likely as Caucasian defendants (12.3%) to be incorrectly classified as high-risk when they did not reoffend. This 12.4 percentage point disparity represents a fundamental fairness violation where one racial group bears disproportionate burden of erroneous high-risk predictions.

The statistical parity difference of -0.143 further confirms systematic bias, indicating that protected group membership significantly influences prediction outcomes. The disparate impact ratio of 0.676 falls well below the 0.8 threshold for adverse impact, strengthening evidence of discriminatory effects. While overall accuracy metrics appeared reasonable, they masked these critical group-wise disparities that have profound real-world consequences for affected individuals.

Remediation Recommendations

Immediate implementation of bias mitigation techniques is essential. Our demonstration of reweighing preprocessing reduced statistical parity difference to 0.003, proving that algorithmic interventions can effectively address some disparities. We recommend deploying multiple complementary strategies: adversarial debiasing during training, threshold optimization for different demographic groups, and regular fairness audits throughout the model lifecycle.

Organizations should establish comprehensive monitoring frameworks tracking false positive rates, demographic parity, and equalized odds metrics across racial groups. Transparent documentation of model limitations and observed biases must accompany any deployment. Ultimately, while technical solutions can mitigate some disparities, careful consideration of the societal context and potential harms remains crucial for ethical deployment of recidivism prediction systems.

Conclusion

The COMPAS system exhibits measurable racial bias that translates into unfair outcomes for African-American defendants. Addressing these issues requires both technical interventions and thoughtful policy considerations to ensure automated risk assessment tools promote justice rather than perpetuate existing disparities