Algorithmic Fairness in Criminal Justice

The COMPAS Case

Rafael Grazzini Placucci

December 14, 2025

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

- Designed to predict a defendant's risk of recidivism based on factors such as criminal history, age, etc.
- Widely deployed across the US to inform bail, sentencing, and parole decisions.
- In 2016, ProPublica found evidence that COMPAS disproportionately misclassified Black defendants as high-risk.

Recidivism

The tendency of a convicted criminal to reoffend.

The goal of this project is to examine the fairness of the COMPAS algorithm by comparing **risk score distributions** and **false positive rates** across race groups.

ProPublica Dataset

We perform this analysis using a dataset of **7,214 criminal defendants** from Broward County, Florida, scored by the COMPAS algorithm between 2013 and 2014.

For each defendant (row), we observe:

- `race`: The race of the defendant (White, Black, Asian, etc.)
- `decile_score`: The predicted risk of recidivism, from 1 to 10
- `two_year_recid`: Whether the defendant was re-arrested within two years

Other metrics in the dataset were not used in this analysis.

Distribution of Risk Score Across Race Groups

Let $S = \texttt{decile\_score} \in \{1, \ldots, 10\}$ denote the COMPAS risk score and $G = \texttt{race} \in \{\mathrm{White}, \mathrm{Black}, \ldots\}$ denote the race group.

We examine whether COMPAS assigns systematically different risk scores to defendants across race groups by considering:

1. The mean score by race:

$$\mu_g = \mathbb{E}[S \mid G = g]$$

2. The inter-quartile range of score by race:

$$\mathrm{IQR}_g = Q_{0.75}(S \mid G = g) - Q_{0.25}(S \mid G = g)$$

We use *bootstrapping* to estimate the 95% confidence intervals around these statistics.

Examining Racial Bias in Risk Scores

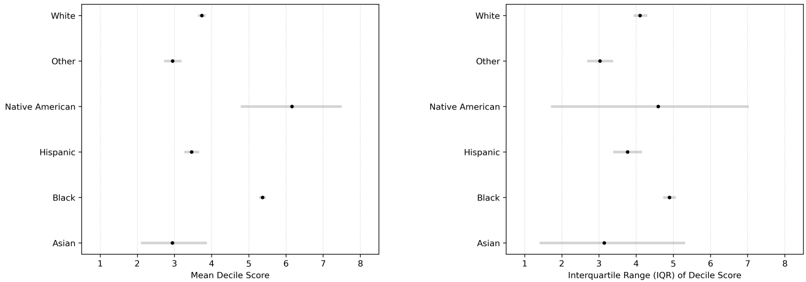

Figure 1: Left: Forest plot of mean risk scores by race group. Right: Forest plot of IQR of risk scores by race group. The results indicate that Black defendants are systematically assigned higher and broader risk scores than White defendants.

False Positive Rate Disparity

Formally, we define the false positive rate for a race group g as:

$$\mathrm{FPR}_g = P(\hat{Y} = 1 \mid Y = 0, G = g)$$

where:

- $G = \texttt{race} \in \{\text{Black}, \text{White}, \text{Asian}, \text{Hispanic}, \ldots\}$
- $Y = \texttt{two\_year\_recid} \in \{0, 1\}$
- $\hat{Y} = \mathbf{1}\{\texttt{decile\_score} \geq t\}$

The *disparity* in FPR between two race groups $j, k \in G$ is then given by:

$$\Delta_{j,k} = \mathrm{FPR}_j - \mathrm{FPR}_{k \neq j}$$

Testing Racial Equality in False Positive Rates

We assess whether COMPAS satisfies equality across race by testing whether FPR differs between two race groups:

$$H_0 : \Delta_{j,k} = 0$$
$$H_A : \Delta_{j,k} \neq 0$$

This tests the claim that defendants of a race group j are more or less likely to be falsely labeled as "high-risk" than defendants of race group k.

We use a *permutation test* that treats race labels as exchangeable under the null of no racial disparity in FPR.

FPR Disparity: Black vs White Defendants

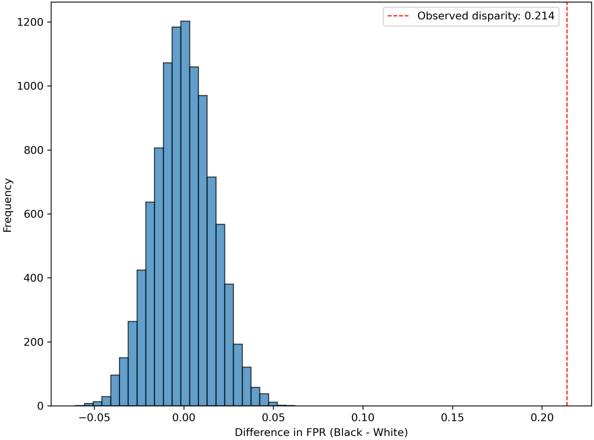

Figure 2: Permutation null distribution of FPR differences between Black and White Defendants, with the observed disparity $\Delta^*_{j,k} = 0.214$ marked in red. We **reject** the null hypothesis of no difference with $p \approx 0$.

Summary

- Black defendants are on average assigned a risk score of 5.37, as opposed to 3.74 for White defendants.
- Black defendants are assigned a wider range of risk scores than White defendants, as shown by an IQR of 4.90 versus 4.10.
- Black defendants are 21.4% more likely to be misclassified as "high-risk" than White Defendants.
- The permutation test indicates that there is strong evidence to suggest that COMPAS is biased against Black defendants.

For more detail, see: github.com/rplacucci/fairness-compas

Implications for Algorithmic Fairness

This analysis raises serious concerns about using algorithmic decision-making in criminal justice, where errors can have severe consequences for defendants' freedom, job prospects, and life outcomes.

Technical sophistication by itself does not ensure fairness. Real accountability requires continuous auditing, clear explanations of tradeoffs between different fairness goals, and public deliberation about whose interests these systems are meant to serve.

References

Angwin, Julia et al. (May 2016). "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks". In: *ProPublica*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 12/14/2025).

Christian, Brian (2020). *The Alignment Problem: Machine Learning and Human Values*. New York: W. W. Norton & Company.

ProPublica (2016). *compas-analysis: Data and analysis for "Machine Bias"*. https://github.com/propublica/compas-analysis. GitHub repository. (Visited on 12/14/2025).

Questions?