

COMPAS Fairness Worksheet

2025-02-26

Your Name Here: _____

0.1 The COMPAS data

Table 1 shows the [confusion matrices](#) for Black and white defendants, respectively, based on the actual COMPAS data.

Table 1: Confusion matrices for COMPAS defendants.

(a) Black defendants				(b) white defendants			
	Low	High	Total		Low	High	Total
Surv	990	805	1795	Surv	1139	349	1488
Recid	532	1369	1901	Recid	461	505	966
Total	1522	2174	3696	Total	1600	854	2454

0.2 Disparate impact

Consider the following mathematical definition of **disparate impact**

$$\frac{\Pr(High|White)}{\Pr(High|Black)} < 1 - \epsilon,$$

where ϵ is a small positive constant, and the algorithm exhibits ϵ -disparate impact if and only if this inequality is true.

- (1) What would it mean for the LHS to equal 1?
- (2) How far from 1 would you tolerate this score deviating before alleging bias? Determine a value for ϵ in your group.
- (3) Compute the disparate impact score using the definition and confusion matrices in Table 1.
- (4) How does this score relate to ProPublica's allegations in Angwin et al. (2016)?

0.3 Equalized Odds

An algorithm exhibits ϵ -**equalized odds** if and only if:

$$|\Pr(High|White \cap Survived) - \Pr(High|Black \cap Survived)| < \epsilon$$

for some $\epsilon > 0$.

(5) Does the COMPAS algorithm exhibit equalized odds (using your value of ϵ)?

(6) Is COMPAS biased according to equalized odds? What values of ϵ would you be willing to tolerate?

0.4 Well-calibration

An algorithm is said to be ϵ -**well-calibrated** if and only if:

$$|\Pr(High|Black) - \Pr(Recidivated|Black)| < \epsilon,$$

for some $\epsilon > 0$ and *for all groups* (e.g., the property holds for white defendants as well).

(7) Is COMPAS well-calibrated? What, if any, rebuttal would you make against ProPublica's claims of bias?

References

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." ProPublica; ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.