COMPAS Fairness Worksheet Part 2

2025-03-05

Your Name Here: _____

0.1 The COMPAS data

Recall Table 1 which 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

	Low	High	Total
Surv	990	805	1795
Recid	532	1369	1901
Total	1522	2174	3696

(b) white defendants

	Low	High	Total
Surv	1139	349	1488
Recid	461	505	966
Total	1600	854	2454

0.2 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).

(1) Recall that we found the following calibration scores:

$$|\Pr(High|Black) - \Pr(Recidivated|Black)| = \left|\frac{2174}{3696} - \frac{1901}{3696}\right| = |0.5882 - 0.5143| = 0.0739$$

$$|\Pr(High|White) - \Pr(Recidivated|White)| = \left|\frac{854}{2454} - \frac{966}{2454}\right| = |0.3480 - 0.3936| = 0.0456$$

(2) Given these calculations and $\epsilon = 0.1$, is COMPAS well-calibrated?

(3) What is the goal of calibrating an algorithm? What are some consequences of a poorly-calibrated algorithm?

0.3 Balance for Negative Class:

An algorithm is said to have ϵ-balance for the *negative* class if and only if:

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

(4) Compute the two probabilities above to determine if COMPAS displays balance for the negative class.

0.4 Balance for Positive Class:

An algorithm is said to have ϵ-balance for the *positive* class if and only if:

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

(5) Compute the two probabilities above to determine if COMPAS displays balance for the positive class.

(6) Balance for the positive and negative classes have similar computations but imply different protections for different groups. What are some of these implications?

(7) Does COMPAS meet all three statistical criteria for fairness (i.e., well-calibration, balance for the negative class, and balance for the positive class)?

0.5 Philosophical Approach

(8) You subscribe to the _____ branch of ethics. From this perspective, which statistical criteria for fairness would you prioritize the most? Which criteria would you deprioritize?