HW 5

Riley Richardson

12/29/2023

This homework is meant to give you practice in creating and defending a position with both statistical and philosophical evidence. We have now extensively talked about the COMPAS [1] *data set, the flaws in applying it but also its potential upside if its shortcomings can be overlooked. We have also spent time in class verbally assessing positions both for an against applying this data set in real life. In no more than two pages* [2] *take the persona of a statistical consultant advising a judge as to whether they should include the results of the COMPAS algorithm in their decision making process for granting parole. First clearly articulate your position (whether the algorithm should be used or not) and then defend said position using both statistical and philosophical evidence. Your paper will be grade both on the merits of its persuasive appeal but also the applicability of the statistical and philosophical evidence cited.*

Using the COMPAS algorithm is not justifiable because (1) any utilitarian benefit is outweighed by the probability of innocent people being made to suffer and (2) that suffering will not be experienced fairly across demographic lines.

The first question we must ask is precisely whom we owe consideration. I argue that the basis of ethical consideration is sentience: if someone or something is capable of suffering, we have obligations to that person or thing. Though one distinction I'd like to make is on the *scale* of those obligations. Most of the benefits of COMPAS are generally abstract — denying parole to a "high risk" subject creates the potential that someone, somewhere, at some point will suffer due their actions, but we cannot say who, where, or when. Meanwhile, most of the detriments are concrete — we are talking about subjecting a specific, known person to measurable suffering.

The deontologist would argue that these consequences shouldn't be treated with the same weight. Intending to harm someone directly is much different than initiating the *potential* that someone comes to harm through one's inaction. The utilitarian, however, would argue that passively allowing people to come to harm is the same as actively harming someone. I do not contest the utilitarian argument in favor of the COMPAS algorithm — *accurately* predicting recidivism, especially violent recidivism, would certainly have a positive impact on society — but I will endeavor to show that any benefit from the program's accurate classifications is outweighed by its potential and tendency to misclassify.

Table 1: COMPAS Performance

	Non-Violent		Violent	
	FPR (non-violent)	FNR (non-violent)	FPR (violent)	FNR (violent)
All	32.35 %	37.40 %	27.93 %	47.15 %
Black	44.85 %	27.99 %	38.14 %	38.37 %
White	23.45 %	47.72 %	18.46 %	62.62 %

Table 1 shows the rates at which COMPAS returns false-positive and false-negative predictions for violent and non-violent crimes in percentages.[3] This essentially tells us who will suffer because of your mistake:

[1] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

[2] knit to a pdf to ensure page count

[3] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

if a future recidivist is misclassified as having low risk (a false negative), the misclassification will afflict the general public; if a non-recidivist, however, is misclassified as having a high risk (a false positive), that innocent (for the 2nd predicted crime) person will be afflicted.

We can immediately see that COMPAS is more likely to falsely classify Black subjects as “high risk” than white subjects — nearly twice as likely for non-violent crime; more than twice as likely for violent crime. This false positive rate has a substantial impact: 44.85% of Black people classified as having a “high risk” of committing non-violent crimes did not go on to commit crimes. In other words, nearly half of the Black people you deny parole using this program could be innocent in practice, compared to under a quarter of white people.

Two statistical measures of fairness we can employ here are (1) predictive equality — the difference between false positive rates — and (2) equal opportunity — the difference between false negative rates. We could also use disparate impact and statistical parity, but these do not consider the underlying disparity between recidivism rates, so the afformentioned two are the most appropriate in this case.

The difference between the false positive rates of Black and white subjects is 21.4% for non-violent recidivism and 19.68% for violent recidivism. The legally accepted threshold is 20%[4]; therefore, COMPAS’ performance on non-violent recidivism fails the predictive equality test (though just barely). The difference between the false negative rates of Black and white subjects is 19.73% for non-violent crime and 24.25% for violent crime. This time, COMPAS fails the equal opportunity test for non-violent crime and passes for violent crime.

Not only does COMPAS partialy fail each of these fairness tests, both places the algorithm passes the test are within half a percent of failure. In a criminal trial, the prosecution’s burden is to prove the defendant’s guilt beyond a reasonable doubt — 0.5% is a reasonable doubt. Moreover, COMPAS’ methods are not transparent and its results are not reproducible. We may know the general reliability of the methods used, but it would be impossible to show that those methods were reliably applied in a specific case.

COMPAS only barely passes two statistical fainess tests, and partially fails them, COMPAS’ methods are not transparent, and COMPAS’ weaknesses disproportionately and systematically punish Black defendents. Implementing this algorithm will necessarily cause innocent people to suffer; it is unethical and unfair.

[4] Dana Pessach and Erez Shmueli, “A Review on Fairness in Machine Learning,” *ACM Computing Services* (55), no. 3, p. 51:4