HW 5

Mason Boyles

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 philosohpical evidence cited.

Despite being very tempting, it is evident that a judge using the COMPAS algorithm to supplement their decisions is not advisable. My explanation for this position can be divided into two main sections. The first will be a statistical critique of the COMPAS algorithm, and the second will be a philosophical discussion about its' shortcomings as well as the implications it has.

From the perspective of a statistical critique, the main issues that I see with the COMPAS algorithm are that it doesn't pass important fairness criteria and that it is completely black box. For the first fact, it doesn't satisfy the important fairness criteria of equalized odds. This means that among the binary of black or white, the rate of false positives should be roughly equal, it shouldn't be overly harsh on one race, but more lenient on another race. We can confirm this by subtracting the false positive rate for white people from the false positive rate of black people.

$$P[\hat{y} = 1|(S = 1 \cap y = 0)] - P[\hat{y} = 1|(S \neq 1 \cap y = 0)]$$

The result comes out to .218, which is greater than the conventional cutoff of .2. This is statistical evidence that proves that the COMPAS algorithm is more likely to overestimate that black people will be repeat offenders as opposed to white people. The second statistical issue I have with the COMPAS algorithm is that it is completely black box. This means that we have no idea how exactly it comes to its decisions, the only things we know are who the input is and it's final decision. This is an issue because it makes auditing its' decision harder since we have no idea what exact factors led to it making a decision. It would be more helpful if it could provide some insights about the person that is inputted and then allow the judge to make a decision based on these insights.

On the other hand, from the perspective of a philosophical critique, the use of COMPAS goes against virtue ethics and it also poses more potential risk than benefit. From the perspective of morality, it is important to make choices that are in line with key virtues. In this case, the virtues of fairness and transparency are both violated. This is because of the reasons I listed in my statistical critique, it is unfair to incorrectly classify black people as repeat offenders at a significantly higher rate than white people, and it is not transparent to make the algorithm completely black box because this makes auditing harder. In addition, from a utilitarian perspective, the use of COMPAS provides more potential risk than benefit. The primary benefit is that it has the potential to increase the accuracy of classification by a judge. However, I would limit this by stating that COMPAS was less than 70% accurate, which is likely pretty comparable or lower than the accuracy of a

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

[2] knit to a pdf to ensure page count

judge anyways. Further, cases where judges struggle to make decisions are probably generally the same cases that COMPAS struggles to make decisions. On the other hand, the risks of COMPAS are the potential for it to contribute to racial biases and the potential for corruption. It can contribute to racial biases for reasons that I have already mentioned, so I will go into detail about how it could lead to corruption. One way it could lead to corruption is that it cannot be audited for integrity. From the perspective of an outsider, a judge can just make the decision based on whatever they want and then instead of providing full justification, it could be backed by the algorithm. It provides a lazy backup solution for judges who want a quick answer and reassigns the blame to an algorithm. Having the blame completely on the shoulders of a judge provides more incentive for them to make a just and thorough decision and this would offload some of that responsibility. Along with this, it feels less trustworthy for a decision to be made based on a process unbeknownst to you than it does to have it done by a judge who can be asked to explain his reasoning.

All in all, it is clear that the COMPAS algorithm is not ready to be used by judges for recidivism of convicted felons. The reason I word it this way is because I believe in the idea, but acknowledge that it would require some revision. These theoretical improvements include better accuracy, equalized odds, and transparency about methods and insights that contributed to its' decision.