COMPAS/ProPublica Data Analysis Project
Science of Data Ethics (CIS 399)
Prof Michael Kearns and Dr Kristian Lum
Spring 2020

Due Date: Monday March 23, details TBD. It's likely that we'll just ask you to stitch your writeup into a single PDF and submit via email.

Some added guidance/specifications for your projects and their writeups:

- You should view your project as an opportunity to demonstrate what you've learned about algorithmic fairness and how it interacts with machine learning and predictive modeling. Try to be scholarly --- i.e. if you mention something related to something in lecture or one of the readings, cite it specifically so it's clear what aspect of class it relates to.

- Your project should contain significant data analysis and/or modeling on the COMPAS data set.

- Your writeup should be exceedingly clear on what your precise methodology was --- exactly what analyses you performed, how you processed or modified the data, which variables or fields you looked at, and filtering or normalization you did, etc. If you wrote code or scripts, you should include them in your writeup. Imagine that the goal is your writeup contains sufficient detail for others to try to replicate your findings. If the methodological details are long, put them in a technical appendix.

- Any plots, charts or figures you include should be self-contained and clear --- axes should be clearly labeled, the units used should be specified etc. Figures should be large enough to be legible but not excessively magnified.

- The above notwithstanding, your writeup should ***not*** simply be a "pile of charts and code". It should equally contain thoughtful narrative and commentary on what your findings "mean", their potential implications for policy or practice, your own opinions on what you found, etc.

- As a very rough guide, I'm imagining most writeups will be in the approximate neighborhood of about 20 pages, including figures. As mentioned previously, more will be expected of larger groups, both in terms of content and thought.

Below are some potential project ideas. They are meant only to be suggestive of the kinds of things you could do, and you are encouraged to be creative and original. Also, these ideas range

from the very broad to the very specific, so not all of them are ambitious enough to be full-fledged project maps. Whatever project you choose, it should be grounded in actual analysis of the dataset provided. As mentioned in class, your project should contain both a strong quantitative/analytical component, and also a strong qualitative/narrative component.

If you search around a bit, you will find some open-source packages for implementing various forms of fair machine learning; one example is https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/. Another is here: https://github.com/fairlearn/fairlearn, which implements both the "bolt-on" approach of the Hardt et al. paper as well as an "in-processing" approach.

- Use the COMPAS risk scores in the dataset to analyze the tradeoffs between error and the various fairness measures we've considered in class, such as false positive/negative rates, positive/negative predictive value, etc. Consider the Pareto curves for each tradeoff and discuss policy/legal implications.

- Consider fairness by other attributes, such as gender and age.

- In class we've discussed potential tensions/tradeoffs between different fairness notions, and also between e.g. racial and gender fairness. Is there evidence for such "fairness fighting fairness" in the data?

- Use machine learning to build your own risk assessment model, either with or without the COMPAS score itself as an input to your model. Can you build models that "beat" COMPAS, in the sense of having a better (more "southwest") Pareto curve? What features are most important in your model?

- Compare different types of ML models (e.g. regression, decision trees, boosting, neural networks, support vector machines, etc.) and the various fairness/accuracy tradeoffs they present

- Can any of the text fields in the dataset (e.g. defendant names, descriptions of current incarceration charge/crime) help in building better or more fair risk assessment models?

- For any particular measure of fairness or discrimination --- e.g. the difference in false incarcerations of whites vs. blacks --- there are two ways we would reduce this difference. Assuming that blacks have the higher false incarceration rate, we could lower theirs towards that of whites; or we could raise the false incarceration rate of whites towards that of blacks. If we agree that the former would be better than the latter, is this actually what more fair models do?

- As Prof Berk pointed out in his guest lecture, the various fairness metrics we have been considering are all about fairness to criminal defendants, and make no mention of

fairness to victims or the costs to society of recidivism. An interesting and open-ended idea is to try to say something about this issue that is driven by the dataset.

- Prof Berk reported that on different datasets, simply training a model on only the white populations resulted in a model that gives roughly the same confusion matrix when applied to blacks or whites, and thus is approximately fair by the measures we've been considering. It would be interesting to see if one could carefully replicate their methodology and finding (or not) his finding on our dataset.

- One can think of the analyses shown in class so far as "post-processing" approaches to fairness: we first build a risk assessment model that ignores fairness entirely (as in the linear regression we examined) or one that was trained with a different fairness notion in mind (as with COMPAS), and then we try to enforce fairness on top of this model by adjusting the threshold(s) for recidivism prediction. A natural alternative is "in-processing", where we seek to actually embed our fairness notion in the training process itself. An interesting project idea would be to investigate in- vs. post-processing.

- Re-define recidivism by changing all instances of recidivism where the person was charged with something minor from a 1 (did recidivate) to a 0 (did not). See how the Pareto curves look for a model trained on this new outcome.

- For people who are not into altering the objective function, one could look at the role of variable selection in fairness. Is there some set of covariates (say, if you are restricted to just using prior_count) that looks different in terms of whatever fairness metrics you pick for models built using standard fitting functions? There are few enough covariates that one could just try all possible combinations of covariates to see how each model performs.

- In addition to looking at costs to society of crimes committed, one could look at costs of detention both in terms of the cost to incarcerate per day as well as lost income for person's family, etc.