CS 584: Final Project

Video Due: May 8, 2022, by 5:00 PM (link posted on Piazza)
Project Report Due: May 12, 2022, by 5:00 PM on Gradescope

- You **must** work in a team for this project. Teams can consist of either 2 or 3 members, and all members must be enrolled in the same section. Post on Piazza to start or join a team. Individuals submitting projects will receive a deduction of 15 points from their project grade. The grade for the project and the video will be shared among all team members, and there is no advantage from doing the project alone or in a smaller team.
- The final deliverable is twofold:
  1. A (maximum) 4-page PDF file prepared using the single column ACM submission template (Word or LaTeX) available at `http://www.acm.org/publications/authors/submissions`. Please make sure you are using the (single-column) **submission** template (you can get rid of extraneous information on the first page). The direct link for Word is `https://www.acm.org/binaries/content/assets/publications/taps/acm_submission_template.docx` and the direct link for LaTeX is `https://www.acm.org/binaries/content/assets/publications/consolidated-tex-templat` `acmart-primary.zip`. Note that your work will be judged on the quality of the writeup you submit! While we may ask to see your code in case of questions, you are not expected to submit it unless requested. Therefore **the quality of the writeup becomes even more important.**
  2. A (maximum) 5-minute video presentation of your project. Post a link to this video (which you should keep accessible through at least May 31) to Piazza, public to the class, by the deadline. You must also include the link at the end of your project report submission.
- Each group should submit only **once** on Gradescope. There is a place where you can put in other team members. Multiple submissions by members of the same group will be penalized.
- The project is not intended to be huge – around 2 homeworks worth of work, giving you the option to either be creative or to use what you've learned in a more complex context.

Project Description

We are providing two "readymade" project ideas, and the third option is to design your own project, with a little bit of guidance. Each of these three options is described in its own section below.

1 Option 1: Auditing for Bias

This option is based on the ProPublica COMPAS dataset, but you could find an alternative dataset as well if you are interested, and work on a very similar project. For the COMPAS dataset, the initial task is to predict whether a defendant in a criminal case will recidivate (be arrested and charged for another crime, labeled as 1, as opposed to 0) within the next two years. First, look at the datasets available at `https://github.com/propublica/compas-analysis/`. You will see that there are several variants of the data, and you should feel free to use any (or several). Split your dataset up into training and test sets (keeping 1/3 for testing), and then, based on only the training set, find a model that you would like to use, given that it will be evaluated on the test data, based on any standard measure that you care about (AUC, accuracy, or F-1 score, for example). You should perform and document all the usual steps for model selection. This could include feature selection or engineering as well as choosing which type of model and which hyperparameters to use. Remember not to peek at your test data. Once you are done with model selection, you should then check the performance according to your pre-selected criterion on the test portion of your dataset.

You are now going to consider whether the model is fair. In the true labels, As a possible application, suppose that your predictions are going to be used to decide which individuals to release on bail or parole (depending on whether they are awaiting trial or serving a sentence).

The first question we will ask is whether there is bias in terms of opportunity cost, by comparing false positive rates. To do so, compute the probability that an individual is predicted positive by your algorithm given that they BOTH (1) did not actually recidivate, and (2) are categorized as African-American by the race variable. Compare this to the probability that an individual is predicted positive by your algorithm given that they BOTH (1) did not actually recidivate, and (2) are categorized as Caucasian by the race variable. Report both the numerators and the denominators for your calculations, in addition to the probabilities. Based on this, write one or two paragraphs about whether you think the algorithm is biased towards one or the other group and the potential societal implications of using such an algorithm.

The second question we will ask is whether there is bias in a different sense, by comparing calibration. To do so, compute the probability that an individual recidivates given that they BOTH (1) are predicted positive by your algorithm, and (2) are categorized as African-American by the race variable. Compare this to the probability that an individual recidivates given that they BOTH (1) are predicted positive by your algorithm, and (2) are categorized as Caucasian by the race variable. Report both the numerators and the denominators for your calculations, in addition to the probabilities. Based on this, write one or two paragraphs about whether you think the algorithm is biased towards one or the other group based on this second measure and the potential societal implications of using such an algorithm.

Now write about your insights based on comparing results along these two metrics. Which measure do you think is more appropriate in this domain, or for different groups who care about the domain? Can you think of another domain where the opposite conclusion would hold?

Next, consider the role of the "race" variable as a protected feature, which means that one may not discriminate on the basis of these features, and it is important to ensure that outcomes are fair with respect to them. If your initial model used "race" how do the results change when you explicitly remove that feature? If your model didn't use it, how do the results change when you do? What does this tell you about the direct use of protected features?

Finally, try the whole experiment using a classifier designed to be more fair, for example the classifiers that attempt to achieve demographic parity or equal opportunity described here: `https://scikit-lego.readthedocs.io/en/latest/fairness.html`. Think about what

kind of fairness you want (e.g. parity of false positive rates) and what methods might obtain that. What do you observe now? Is there a tradeoff between accuracy/AUC/F-measure and the fairness objective?

2 Option 2: Cost sensitive learning

An overview of this problem is the following. You have access to a dataset with 481 features and **two** target variables. There is a separate learning set (the training set), and validation set. The task is to solicit donations for an organization. The fields are information about people who were sent a mailed solicitation to donate, and the target variables are (1) whether or not they donated, and (2) how much they donated. Each solicitation costs 68 cents to mail, and the average donation over the solicited population was 79 cents. However, only about 5% of the population donates. You goal is to maximize the difference between the amount donated and the costs of mailing on the **validation set** provided. Thus, you have to decided whom to solicit – if you do not solicit, there is no cost, but also no reward from that individual. If you do solicit, the gain is the amount they donate minus 68 cents. The datasets and a lot of anciliary information, are available at `https://kdd.org/kdd-cup/view/kdd-cup-1998/Data`.

You can be creative in approaching this problem. In order to solve this problem, you will need to try and identify those who are likely to donate, but also be very sure not to miss out on soliciting those who give high value donations. Be sure to read the files provided in some detail before you start, so you understand the difficulties. At the very least you should be implementing a method that provides scores or rankings (not just predictions) so you can fine tune the threshold(s) that you use. You may also want to experiment with predicting either the "high value" donors, and/or using a cost-sensitive approach to classification, and/or using a regression technique in addition to a classification technique. Beware that the small number of positive examples can be problematic for some classification techniques, which is why a scoring method is particularly important.

You should design your method entirely on the learning (training) dataset before performing just a final test on the validation dataset. In your report, you must describe your overall approach to the problem, validation approaches you used on the training dataset, how you expected it to perform on the validation set, as well as your final performance on the validation set provided. While there is one final evaluation metric (profit), you should also report other metrics (like accuracy, precision, recall, and any others that quantify high value donors or donations being captured or missed).

You are going to be judged on the way you describe your approach and how sensible it was, not on the outcome, so please don't "cheat" by peeking at the validation set before having completed the design of your method.

3 Option 3: Your own idea

Come up with an interesting *question* and an interesting *dataset*. You may foreground either the question or the dataset, but it must be a question that can be reasonably asked using that data.

For datasets, feel free to browse Kaggle (`https://www.kaggle.com/`) or the UCI ML repository (`https://archive.ics.uci.edu/ml/index.php`). However, you may also collect a dataset of interest to you and make that the primary focus of the project (that is, you could apply the tools we've learned to answer a standard question on the novel dataset). Here are some examples of questions you might be interesed in answering. Remember that these are just ideas, and you should feel free to modify them or come up with your own.

- Can you collect a novel dataset related to a topic of current interest (e.g. elections, Covid, something else) using the Twitter API and then answer an interesting question (e.g. how different countries / areas / states respond to something?)
- How do different classifiers compare in terms of different fairness metrics on different datasets with sensitive or protected features like race or gender?

Please post your idea to Piazza as a private note by April 19 (at the latest), with thoughts on both the question and the data you're interested in, and we will try and give you feedback on the idea ASAP.