Pa.06 – Case Study

Resources

- Google Colab: "Case Study.ipynb"

Introduction

In this final week, we'll bring together what we learned in the previous weeks to look at one larger case study. We'll work with semi-real data to construct a case that's similar to a case of algorithmic unfairness that's already been in the news and that you may remember from the very first week (Austria's employment agency).

Assume that you are responsible for an employment agency in Pennsylvania (PA) in the US. The employment agency pays for people's unemployment benefits and is tasked with helping them find work. There are programs for recently unemployed people that work fairly well, but at the moment there is no good program for longtime unemployed people. However, there is a team of specialized job coaches that consults longtime unemployed people in Pennsylvania and supports them in writing job applications, finding possible employers, etc. These job coaches have a promising success rate: They have helped 66% of longtime unemployed people who they have worked with find a job. Of course, participating in the job coaches' program comes with a fee. As an agency financed by taxes, you have the responsibility to use your financial resources efficiently. The "business case" is easy: Each year in which a person is unemployed costs the taxpayers CHF 30k on average. The job coaches' program takes one year and costs CHF 20k per person. If it is a success and the longtime unemployed person finds a job in the first year and thus does not collect unemployment benefits in the years after that.

You are aware of the historical and systemic discrimination faced by the black community. Therefore, you and the public are concerned about the fairness of how you assign access to the job coaches among black and white longtime unemployed people.

In this exercise sheet, you will have to evaluate the current model in terms of monetary efficiency and fairness and iteratively improve on it. Along the way, you will have to justify your choices to members of the public who care about fairness and about how you spend the tax money. Some tasks include instructions such as "Justify your choice to the public." This asks you to defend your choices to the public (imagine investigative journalists or interested members of civil society). Explain why you made these choices and try to convince the public that they are sensible.

Setup

The dataset we will use was constructed from US census data for the purpose of testing algorithmic fairness metrics and mitigation techniques. This specific dataset has to do with the employment. You can find a detailed explanation of the dataset's features in Appendix B.4 of the paper that constructed the dataset.

N.B.: For our use case, we relabel the target variable from its original meaning to "this person benefits from working with the job coaches". In our use case, we will thus not assume that the dataset is a representative sample of the full population (which it really is) but rather of longtime unemployed people who have worked with the job coaches. This relabeling of course would not make sense in practice, but we do it here to construct a semi-real dataset to work with for this task.

Y=1 thus means that the jobseeker found new employment after 1 year of working with the job coaches.

Y=0 means the jobseeker did not find new employment despite the help of the job coaches.

The other features are (in this order):

- AGEP (Age)
- SCHL (Educational attainment)
- MAR (Marital status)
- RELP (Relationship)
- DIS (Disability recode)
- ESP (Employment status of parents)
- CIT (Citizenship status)
- MIG (Mobility status, lived here 1 year ago)
- MIL (Military service)
- ANC (Ancestry recode)
- NATIVITY (Nativity)
- DEAR (Hearing difficulty)
- DEYE (Vision difficulty)
- DREM (Cognitive difficulty)
- SEX (Sex)
- RAC1P (Recoded detailed race code).

When constructing the dataset, the authors already applied two filters:

- AGEP (Age) must be greater than 16 and less than 90.
- PWGTP (Person weight) must be greater than or equal to 1.

We further apply the following two filters for our case:

- AGEP (Age) must be greater than 24 and less than 66.
- RAC1P (Recoded detailed race code) must be 1 (White alone) or 2 (Black or African American alone).

The library "folkstables", which was published with the paper that constructed the dataset, gives us access to this data for the years 2014 to 2018 for all states in the US.

Assume that the data we have was collected by the job coaches working with other employment agencies in Pennsylvania. A team of data scientists in your employment agency used said data to train a prediction model on data from 2014. Using our previously stated assumptions, the model predicts the chances that a longtime unemployed person finds a job within the year that they work with the job coaches. To train the model, they first split the data into a training set (80% of the data) and testing set (20% of the data). Then, they built a machine learning pipeline that standardizes the features and then they fit a logistic regression to it. The logistic regression outputs a score between 0 and 1 (model.predict_proba() function). These scores are used to choose between:
D=1: The longtime unemployed person gets access to the job coaching program.
D=0: The longtime unemployed person does not get access to the job coaching program.

1. How good is the trained model?

We now want to evaluate the performance of the model on the testing data.

- As a preliminary test, calculate the accuracy of the model, using the standard threshold of 0.5 for the score (model.predict() function).
- Check the calibration of the model. Hint: Split the scores into 10 bins {[0,0.1),[0.1,0.2),…,[0.8,0.9),[0.9,1]}. For each bin i=1,…,10 , calculate
    - The average score of all individuals in this bin: x_i
    - the average reemployment rate of all individuals in this bin: y_i

  Plot a calibration plot with these (x_i, y_i)-pairs, showing the expected reemployment rate and the actual reemployment rate where the actual reemployment rate also shows the 95% confidence interval. Add a line to the calibration plot that shows what perfect calibration would look like. Write a summary about your findings on the calibration of the prediction model.

2. How well does the model work for the two groups?

Now let's see how well the model works for the two groups of black and white job seekers. We'll again assume that to make binary decisions, we will use the standard threshold of 0.5 (model.predict() function).

- Plot the score distribution for both groups. Describe and interpret the results.
- Is calibration-between-groups achieved?
    - Check for every score bin ({[0,0.1), [0,0.2),…,[0.9,1]} whether there's a statistically significant difference in the actual reemployment rates for the two groups. Note: The groups might have a different expected reemployment rate (i.e., different average scores for the same bin). To simplify this comparison, we'll ignore this potential difference here and just compare actual reemployment rates for each bin.
    - Plot the calibration separately for black and white job seekers. The calibration plot should show the expected reemployment rates on the x-axis (different for the two groups) and the actual reemployment rates on the y-axis (also different for the two groups). Also show the 95% confidence intervals for the actual reemployment rates for both groups. Add a line that shows what perfect calibration would look like. Describe and interpret the results.
- Compare the following metrics for both groups: base rate (BR), positive rate (PR), true positive rate (TPR), false negative rate (FNR), false positive rate (FPR) and true negative rate (TNR). Is there a statistically significant difference for the metrics between the groups? Describe and interpret the results.

3. Utility of the decision maker

As the unemployment agency, you have to be careful about how you spend taxpayers' money: One of your goals is to reduce the amount of money that you spend.

Based on this goal, what's the utility matrix of the decision maker in monetary terms for the two years after each decision? Explain your choice to the public.

Decision maker utility matrix	Y=0	Y=1
D=0		
D=1		

What would be the single threshold that maximizes the decision maker's utility?

- What should it be theoretically, assuming that the predicted probability was equal to the true probability?
- What is the one that you find to maximize the total utility on the testing data, testing the thresholds {0,0.01,0.02,…0.98,0.99,1}? What is the total utility for this threshold?

4. Moral analysis

As a government agency, you (and the people who you serve) don't only care about monetary values though – you also want to be fair to decision subjects. You therefore define a fairness criterion that you use to evaluate how fair your decision-making system is. For this, you follow the pattern of defining a utility matrix V, the sensitive attribute A and the justifier J with the relevant value j. You'll have to carefully choose these variables based on the application context.

V: What's the utility matrix of the decision subjects? Justify your choice to the public.

Decision subject utility matrix	Y=0	Y=1
D=0		
D=1		

A: The relevant groups are given in the dataset: Black and white job seekers.

J: Is there a justifier J that (from the decision subject perspective) justifies differences in the utilities the individuals receive from the decision-making system (i.e., the process that distributes the ??)? Justify your choice to the public.

Fairness criterion:

- Write down the resulting fairness criterion as an equation and simplify it as much as possible. Is your fairness criterion one of the well-known criteria derived from the decision matrix?
- Is the fairness criterion fulfilled under the experimentally best decision rule from task 3? Measure the metric for both groups and check for statistical significance.

5. Trade-off between DM utility and DS fairness

Use the FairnessLab to create a Pareto front and pick (group-specific) thresholds that you could justify to the public.
Here are the steps you should take:

- Create a dataset that fulfills the requirements of the FairnessLab (check the FAQ section of the FairnessLab).
- Feed this dataset to the FairnessLab and configure it, so that it represents your DM utility matrix (from task 3) and your fairness score (from task 4).
- Check the Pareto plot for 101 thresholds per group ("How many thresholds do you want to test for each group?") and explore different points (i.e., threshold combinations) in the plot. Choose one that offers a good trade-off in your opinion.
- What threshold did you end up choosing? Justify your choice to the public.

6. Deployment four years later

The model is now deployed in the same state for the next four years. We now want to evaluate how well the model does at that point (so four years later).
Go through tasks 1 through 5 again but use the data from 2018 (full dataset, without train-test-split) to evaluate the model from 2014. For tasks 3 and 4, the DM utility matrix and fairness criterion will stay the same. What you should reevaluate here is the optimal threshold for the DM on the 2018 dataset and whether the fairness criterion is fulfilled. Think about why things changed if they changed.
You can of course copy and paste your code from tasks 1 through 5. What's important here is the reinterpretation of the results.