Fairness Definitions - An Applied Case Study

Juliet Fleischer

November 2024

1 Potential Problems

- feature selection: based on intuition and context but not assesed in a data-driven way with any algortihm or method - no hyperparameter tuning - missing data imputation for the already reduced dataset (without explicitly analysing missing data patterns or testing) - only one imputed data set (computational reasons) - data inherent challenges (given the the imbalance and selection bias data should probably be treated differently)

2 Data and methods

2.1 Data

Where does the data come from? What is context of the data? How many frisks/ searches/ summons/ arrests in total data/ in subgroup?

2.2 Methods

Missing Values currently imputed everything with the mice package and checked whether the distribution for relevant variables remained approximately the same

random forrest classifier; hyperparameters set (no tuning)

2.2.1 feature selection

set of features we choose should resemble the information the officer had at the time of their decision and control for some other variables, that add to the atmosphere of the situation and might influence it's outcome Depending on which target we look at we choose slightly different features - target: frisked the officer has the least information, only visual - target: searched some more info about potential illegal substance, hard object etc - target: summon issued unclear, don't take - target: arrested not necessarly most info, there are ppl that were not searched or frisked but arrested

2.2.2 target selection

potential targets are frisked, searched or arrested With respect to our overall goal to illustrate fairness notions and give an introduction to the topic, we choose "Arrested" as target. The reasons for this are first, that the imbalance of 0-1 conditioned on the PA is okay for arrested. The proportions are 10 percentage points ways, so not very close but it is okay. Also, based on the feature Importance plot, the model relies a lot on one feature but still it is not solely relying on it.

3 Sources of bias or the context of your data

Maybe we have to illustrate a broad categorization of potential sources of bias because this data case shows actually why it might be important to think about sources of bias in your data first before doing other analysis. Our data e.g. suffers from a sort of selection bias which is, so the bias is already engraved in the data. Bias can be i) in the data itself (unfair data, perfect algo, still unfair result) ii) in the algorithm (fair data, unfair algo, unfair result) iii) a mix of both, something in between, and the feedback loop

4 Group/Statistical/mlr3 Fairness Metrics

4.1 race as PA

PA: race group u - people of color group p - not people of color

Confusion matrix based fairness.acc / Overall accuracy equality We care about true positives and true negatives alike. So the people that were correctly predicted as arrested or the people that were correctly predicted as not arrested The probability for a suspect that is predicted arrested to actually have been arrested and the probability of a subject that is predicted not arrested to actually not have been arrested should be the same across groups. In our case, the proportion of correct predictions among the group u is ... while the proportion of correct predictions among group p is .. The absolute difference in accuracy between the two groups is ...

fairness.fpr / False positive error rate balance / predicive equality For punitive situations this should be kept to a minimum, in the discriminated group the innocent have a higher probability of punishment

fairness.tpr/ equivalent to False negative error rate balance / equal opportunity For punitive situation it is good when it's high but not at the cost of a higher fpr. For assistive situation it should be maximized (all the ones that need help, get the help, all the ones that have a weapon get predicted to have one)

fairness.ppv / Predictive parity In punitive tasks you want this to be high, people who are predicted guilty get the punishment so we should be really sure that they are actually guilty. This however, is also influenced by the prevalance in population, so how many guilty people are in population (correct?, satz von bayes?)

fairness.equalized.odds mean of fairness.fpr and fairness.tpr; don't dinstinguish importance of FP and TP so much but average out, so e.g. when we have super unfair fpr but the tpr is okay then the equalized odds will diminish a bit of this super unfair fpr

fairness.fomr should be minimal for assistive, given that the person is predicted of no need for help, whats the prob that person does in fact actually need help given that the person is predicted innocent, whats the prob that the person is actually not innocent and we missed a guilty person

fairness.cv / Group fairness / statistical parity / equal acceptance rate equal proportion of predicted positive across groups (doesnt account for imbalanced groups and prevalence i would say)

Tendenzen PA - race mit unprivileged = c(black, black hispanic), privileged = others The algorithms scores better in terms of fairness for the assistive fairness measures, which is unsuitable for the task because it is more of a punitive task. So very roughly speaking, the algorithm is better at not predicting arrested when really not arrested and if predicted not arrested to really be not arrested. It is worth in predicting arrested for the ones that were arrested and if they were arrested to predict they really were. All differences are small though, and in other papers even the largest difference we see in this data is overlooked and the classifier deemed as fair regardless.

Naming error in mlr3 - fairness.ppv is predictive parity, but they say that fairness.pp is predictive party - fairness.pp takes more the role of fairness.equalized.odds but for PPV and NPV instead of FPR and TPR - so a better name for fairness.pp would be fairness.equalized.predictive.outcome or sth like this

Custom: Treatment equality, Conditional use accuracy equality

4.2 sex as PA and arrested as target

First and foremost, we observe imbalance in the PA. The majority of recorded cases are male (93,5 %), so we have the same situation as with race as PA. There is way more data for males than for females and in general the majority class tends to have better predictions for any machine learning algorithm. Second, there is an imbalance in number of cases (arrested suspects) in females and males. This has effect on some of the group fairness metrics ("Nuances paper").

4.2.1 Independence

Independence: Statistical Parity/ Demographic parity (not implemented in mlr3) Focuses on the proportion of predicted cases in each subgroup of the protected attribute. Asks whether the proportion of positive predictions is approximately equal among females and males. This metrics makes especially sense when you don't want there to be a disparity in positive predictions in groups. In the context of our dataset this means that one assumes any disparity in arrests between females and males is due to bias and not due to the fact that females correctly are less often arrested/ not arrested. In the NYPD dataset demographic parity is not given. The proportion of arrested suspects among females is higher than for males. Simply said, to obtain demographic parity the classifier would need to arrest males more often or arrest females less often. This could be achieved by lowering the threshold for males or increasing the threshold for females. This of course assumes that we think it is wrong that females get relatively more often predicted as arrested. – insert visualization –

4.2.2 Separation

fairness.fpr/ False positive error rate balance / predictive equality For each group of the PA we look at the proportion of false positives. False positives are predicted as arrested even though they were not arrested. Our classifier doesn't satisfy predictive equality as the fpr is much lower for males than for females. More the perspective of people that were predicted arrested (but were actually not arrested, makes more sense in our scenario as it is more of a punitive task). Often suitable when we want to prevent innocent people from receiving penalty. ¡-¿ satisfied at same time as fairness.tnr

fairness.fnr / Equality of Opportunity/ Equal Opportunity / False negative error rate balance Perspective of people that were arrested. ¡-¿ fairness.tpr

Score-based Instead of conditioning on the true 0 (or true 1) and counting the number of positive (or negative) responses for each groups, we can condition on the true 0 (ture 1) and get the average prediction scores for each group. This is called Balance for the negative class (or the positive class) and i basically the equivalent to Predicitve equality but using the predicted scores and not the predicted labels.

mlr3 fairness graphs: the predicted probability graph shows that the algorithm predicts higher probability for being arrested for females versus males.

Do we deem higher probability of predicted arrested as unfair against or fair? We chose a limited set of features, that should migitate the information/ situation in which the officer chose to arrest the suspect. A high arrestment rate could now mean that the group with high arrestment rate is the discriminated one because arrestment is bad and they get the bad thing more often. Or it is good because we set arrestment equal to guilty and then this means that the treatment is correct and the group with higher arrestment rates was less often stopped for no reason. And less often stopped for no reason would indicate that they are not discriminated against, because the officers didn't assume they did something illegal based on their looks. Or is this exactly what the fairness metrics tell us?

4.2.3 Sufficiency

fairness.ppv / Predictive parity/ outcome test Not given, the ppv is higher for males than for females, which means the proportion of suspects that were correctly predicted as arrested is higher among males than females. ¡-¿ fdr

fairness.fomr

Score-based metrics: Calibration and well-calibration Calibration tends to be better for males than females, this can partly be influenced by the abundance of data we have for males.

5 Individual/ Similarity-based metrics

Fairness through Unawareness (FTU) This definition sets more of a constraint on the process rather than demanding the results of the classifier to fulfill a certain criterion. It demands the practitioner to not use the sensitive attribute in the decision process. Remove the PA from the training data, also called blinding. Advancement of this is suppression, that not only delets PA during training but also all features that are highly correlated with the PA. How much correlation is enough to discard the feature is up to individual choice, no standard cutoff. For our data this would mean exclude sex during training for FTU. For suppression potentially additional variables such as height and weight (t-test shows significant difference in height and weight between males and females). On the other hand, the weight and height is no real proxy in the scenario for the officer because when there is basically no possible way in which they could have seen the weight but not the sex and the other way around.

Fairness through Awareness Causal Discrimination as outlined in Verma and Rubin is a case of FTA that results from specific choice of the distance metric. For each of the suspects in the test set we create a duplicate with only the sex different. See whether get the same prediction. Other possible similarity measures in the feature space could take into account only a subset of the features, such as age, type of crime, reason for being stopped,...

6 Causal Fairness Metrics

7 Beyond binary classification with binary PA

Multilevel PA (more than two levels) Multiple PAs (more than one PA)

8 Limitations and Discussion

- binary PA or multilevel PA? - unequal groups + imbalance due to data bias (discuss data bias and how it is always essential to regard the context of your data)

9 Further frameworks

just mention Bothman et. al framework?