Seminar Thesis

fairML - a Case study of the SQF dataset

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, January, 12th 2025

Submitted in fulfillment of the requirements for the degree of B. Sc.
Supervised by Dr. Ludwig Bothmann

Abstract

In this study we provide an introduction to the most common fairness definitions, illustrating them with the example of the SQF dataset.

Acknowledgement

Contents

List of Figures

List of Tables

1 Introduction

Das ist die erste Zeile in meinem intro.tex File. lorem ipsum dolor sit amet, consectetur adipiscing elit. This is a text citaiton Verma and Rubin 2018

Residual Unfairness

Preposition 2: For group a the scores of the target population are always strictly higher than of the training population. This means that we will learn a comparatevily low threshold for group a. When we employ the algorithm in the target population, group a member will receive the positive outcome more easily (receive benefit of the doubt) because the thresholds is so low. For gorup b the opposite is true. The scores in the training data are really high compared to the overall population. This means we learn a high threshold for group b. When the system is applied on the whole population it will be harder for a random person from group b to receive the advantage because their threshold is so high. Applied on the SQF data this could translate as follows. First of all, the interpretation shifts. $\hat{Y} = 1$ is no longer desirable and we can interpreate scores as riskscores G_g^E. This means a high thresholds for being classified as $\hat{Y} = 1$ is desirable, a low threshold is undesirable. We assume that officers were more lenient to stop black individuals, which means that the scores (probability of actually having committed crime) in the training population of black people are lower than the scores of the target population of black people. $G_b^{Z=1} \preceq G_b^{T=1}$. This means we will learn a lower threshold for black people*(???). When we apply the algorithm to the target populaiton we will be more likely to classify black people as $\hat{Y} = 1$ because the threshold is so low. White people, on the other hand, were selected more strictly. This means that the scores of white people in the training population are higher than the scores of white people in the target population. $G_w^{Z=1} \succeq G_w^{T=1}$. This means we will learn a high threshold for white people. When we apply the algorithm to the target population we will be less likely to classify white people as $\hat{Y} = 1$ because the threshold is so high. – Still unsrue if this makes sense, if a transfered it correctly.
* Why do we learn a lower threshold for black people. Maybe something like this happens: So when a group is super leniently stopped we will have many truly innocent and few truly guilty. For the other group we have many truly guilty and less truly innocent. When now 80% of truly guilty are classified as guilty in the advantaged group then we would want want 80% of the truly guilty to be correctly labelled as guilty in the disadvantaged group. This would only results in lowering the threshold for the disadvantaged group (so making it easier to predict them as guilty) if we predicted low risk scores for truly guilty people in the disadvantaged group Because for equal opportunity we are only looking at the people who were really guilty. So we are basically saying that the large proportion of truly innocent people in our sample of the disadvantaged leads to lower risk scores even in the truly guilty group of the disadvantaged (like a spill over effect). Only then it would make sense to say that a fairness intervention would compensate by setting lower thresholds for the disadvantaged group. Is this happening?
Chaper 6: Case study on SQF data Their main message is always, bias in, bias out. fairness interventions, done on the trainign data are not enough, if your sample is biased,

your model will be biased (even after fairness interventions). They show this in the following way. The goal is to predict innocence of an individual. Such an ADM could help officers decide who to stop in the first place. The SQF data serves as training data and is naturally censored. The censoring process is that we only observe innocence of a person if they were stopped. But the decision to stop someone could be based on a biased decision policy. So we have our censored training data (SQF data). We know that this training data is not representative of the population of NYC in general defined via location specific variables. Kallus and Zhou use train a logistic regression classifier on the SQF data as is and use post-processing proposed by Hardt et al. to ensure Equal Opportunity or Equalized Odds. They use their a weighing technique (proposed by them and inspired by propensity score matching) to simulate the target population. The fairness intervention in the training population produces group-specific thresholds that are then applied to the target population. They use these fairness-adjusted threshold for the target population and still observe unfairness.
But of course they observe unfairness because the fairness intervention they do is a post-processing step and doesn't modifiy the classifier. What am i not getting here?

Bias in, bias out - an alternative perspective

Rambachan and Roth n.d. take a different perspective on the problem of biased training data than Kallus and Zhou n.d. The mechanism they describe works as follows I think!!?): Black people are more leniently stopped, leading to higher stopping rates in for black people in the training data, meaning more training data for this group. Because we stop black peopel more leniently, we record nany innocent black people in our data. In Kallus and Zhou n.d. this would lead to a lower learned threshold* for black individuals. Applied on the target population this would mean that we would predict too many false positive. The threshold estimated from the training data is so low that we classify to many people as guilty because in the target populations the scores are actually higher and meet the threshold easily. In Rambachan and Roth n.d. they say that by stopping (searching, they actually talk about searching, not stopping) black people so leniently, our sample for black people comes actually pretty close to the target population. In other words, the training data for black people is pretty close to the target data for black people, which means that our classifier will work well on the target population for black people.
To summarise, in Kallus and Zhou n.d. bias against a group results in a less representative sample. In Rambachan and Roth n.d. bias against a group results in a more representative sample.
* first this leads to lower risk scores for black individuals. And then via fairness adjustments (e.g. for equalized odds) this leads to lower thresholds for black individuals.
Theorem 1 The prediction for african americans is weakly decreasing in tau. This means, as tau increases (so racial bias increases), the expected value for Y gets actually lower, so closer to zero, so less often predicted to have a contraband. What is happening? Higher tau means lower searching threshold for african americans. So the data for african americans becomes "more noisy", more and more innocent people come into our sample, so we predict lower risk for african americans. In Rambachan and Roth n.d. paper this translates to a more representative training data for african americans and thus also

better performance on the general population of african americans. In Kallus and Zhou n.d. paper the mechanisms is the same, we also estimate lower risks cores for african americans, but then sth else happens. I think in Kallus we then do a fairness intervention that leads us to setting a LOWER threshold for african americans, meaning we predict them as guilty more easily to achieve the same FPR as in the other group. I think in kallus they first formulate it in the strict way, where the police is so biased against african americans that the stopped african americans are LESS likely to actually have a weapon than the general population. But they relax this setting afterwards.
What happens if we train the logistic classifier (to predict weapon yes no) on the SQF as is (Kallus), don't do a post processing fairness intervention (NO Hardt et. al) and test the classifier on the target population (that is created via the weighing method of Kallus and Zhou)? I think according to Rambachan and Roth n.d. we should observe bias reversal.

A Electronic Appendix

Data, code and illustrations are available in electronic form.

References

Kallus, Nathan and Angela Zhou (n.d.). "Residual Unfairness in Fair Machine Learning from Prejudiced Data". In: ().

Rambachan, Ashesh and Jonathan Roth (n.d.). "Bias In, Bias Out? Evaluating the Folk Wisdom". In: ().

Verma, Sahil and Julia Rubin (May 2018). "Fairness Definitions Explained". In: *Proceedings of the International Workshop on Software Fairness*. Gothenburg Sweden: ACM, pp. 1–7. ISBN: 978-1-4503-5746-3. DOI: 10.1145/3194770.3194776. (Visited on 11/16/2024).

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, January, 12th 2025

Juliet Fleischer