Seminar Thesis

FairML and the SQF dataset

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, February, 12th 2025

Submitted in fulfillment of the requirements for the degree of B. Sc.
Supervised by FairML and the SQF dataset

Abstract

In this study we provide an introduction to the most common fairness definitions and subtleties that come with them. We advocate for tackling fairness in a wholeistic way, taking into account how the data was generated and how it will be used. This will be illustrated by a case study on the Stop, Question, and Frisk data (SQF) from the New York Police Department (NYPD).

Acknowledgement

Contents

List of Figures

List of Tables

1 Introduction

Here will be an introduction of common fairness metrics (group, individual, causal) similar to my presentaiton. Inspired by Verma and Rubin 2018 Caton and Haas 2024 Castelnovo et al. 2022
Problem of Inframarginality Corbett-Davies et al. n.d. "In this example, the incompatibility between threshold policies and classification parity stems from the fact that the risk distributions differ across groups. This general phenomenon is known as the problem of inframarginality in the economics and statistics literature, and has long been known to plague tests of discrimination in human decisions" For our case this would mean the risk of risk of the target (of being arrested, of being searched, of having a weapon) is really not the same for all groups in the true populaiton. This is realistic and is to be assumed. THen Corbett-Davies et al. n.d. argues that group fairness will no lead to individual fairness, and the optimal classifier from a utility maximazation perspective for the individual will not be fair for the group.
In case of SQF, the observed crime rate among african americans is higher (according to official statistics from NYPD). So it would make sense that more african americans are stopped because they have higher risk scores in general. But the higher risk scores for african americams should be questioned in the first place. They are of course not due to the fact that african americans are truly more likely to committ crime in the first place, but they have developed over many centuries of racial discrimination and targeted policing (lower socio-economic status and more repoerted crime rates because more polcie in these regions). So in this dataset we basically have the problem that - risk scores in the ture population are really different (african americans higher crime rate than white people) $\rightarrow$ due to historical bias, no objective truth process - do not know yet whether risk scores (a.k.a. crime rate) is higher for african americans in my sample - could be that the crime rate in sample is distributed in the same way as in the true population (african americans have higher crime rate than white people) - could be that the crime rate in sample is distributed in a different way than in the true population (african americans have lower crime rate than white people) $\rightarrow$this would be the extreme strict effect described in Kallus and Zhou n.d. where the stop decision is so biased that we explicitly target innocent african americans (this is likely not the case)
This ties into the comment in Castelnovo et al. 2022 that Separation is appropriate when the true lable Y is an objective truth. Here at first sight we would say, whether someone has commited a crime or not is an objective truth. But in reality, the fact that someone committed a crime is influenced by historic bias. Then enfrocing statistical parity here would be good ?? No because then this would e.g. lead to many innocent white people being wrongly accused of crime because after all at the present white people commit less crimes than african americans.
The story I actually want to tell in the end is that hey this is what our results show. It is not a super clear picture and maybe not what you expect, but to see the situation here clearly we have to take into account historical bias ("infected" Y) and sampling bias (PoC more easily stopped). Historical bias currently no specific method or idea to show it, sampling bias is addressed in the literature e.g. simulating the target population with weighing method.

Definitions of Fairness in Machine Learning

When one starts to get into the topic of fairness in machine learning, it is easy to get overwhelmed by the sheer amount of definitions and metrics that are out there. In this chapter we try to group them in an intuitive way and motivate them in the hope to bring some clarity to readers. It is helpful to group fairness metrics in the following ways. We can distinguish

- 1) group fairness vs. individual fairness
- 2) observational vs. causality-based criteria Castelnovo et al. 2022

Broadly speaking, group fairness aims to create equality between groups and individual fairness aims to create equality between two individuals within a group. Observational fairness metrics act descriptive and use the observed distribution of the data to assess fairness while causality-based criteria make assumptions about the causal structure of the data and base their notion of fairness on this (so basically observational says, fairness is when I can measure equality from my distribution and causality says fairness is when the cause for my decision is not discriminatory against someone or a group). On the basis of these fundamental ideas, a plethora of formalisations have emerged. Most of them concern themselves with defining fairness for a binary classification task and one binary protected attribute (PA). The extension to a multiclass PA is the easiest. The extention to multiple sensitive attributes, on the other hand, brings challenges with it. Also, the extension from binary classification to other tasks, such as neural networks, LLMs and other models is subject of ongoing research. As this work is meant to help you start thinking about fairness in machine learning, we will limit ourselves to the binary classification case. Specifically, we want to use the following running example, inspired by our case study on real data in chapter x. The crime rates in NYC should be decreased with the help of a new AI tool. Specifically, the administration orders a team of machine learning experts to design an automated decision-making system that should predict criminal activity of a person. It should be employed by police officers to decide whether to stop a person or investigate them further. Past police stops serve as training data. Given the history of racial profiling in the United States, it is reasonable to raise concerns about racial decision patterns the algorithm could learn from. First, we approach this from a group fairness perspective.

Group fairness

The notion of fairness underlying group metrics is that discrimination of certain groups of the population defined via the protected attribute should be prevented. Group fairness can be grouped into three main categories, independence, separation, and sufficiency.
Independence is in a sense the simplest group fairness metric. It requires that the prediction $\hat{Y}$ is independent of the protected attribute A, so $\hat{Y} \perp A$. In other words, the positive prediction ratio (ppr) should be the same for all values of A. For a binary classification task with binary sensitive attribute this can be formalised as demographic parity/statistical parity $P(\hat{Y}|A = a) = P(\hat{Y}|A = b)$. The other two groups of group fairness metrics, Separation and Sufficiency can both be derived from the error matrix. **Separation** requires independence between $\hat{Y}$ and A conditioned on the true label Y, so

$\hat{Y} \perp A|Y$. This means that the focus is on equal error rates between groups, which gives rise to the following list of fairness metrics:

- Equal opportunity/ False negative error rate balance: $P(\hat{Y} = 0|Y = 1, A = a) = P(\hat{Y} = 0|Y = 1, A = b)$ or $P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$ `mlr3`: `fairness.fnr`, `fairness.tpr`
- Predictive equality/ False positive error rate balance: $P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b)$ or $P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$ `mlr3`: `fairness.fpr`, `fairness.tnr`
- Equalized odds: $P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = b) \forall y \in \{0, 1\}$ `mlr3`: `fairness.equalized.odds`
- Overall accuracy equality: $P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = b)$ `mlr3`: `fairness.acc`
- Treatment equality: $\frac{\text{FN}}{\text{FP}}\big|_{A=a} = \frac{\text{FN}}{\text{FP}}\big|_{A=b}$

Equal opportunity requires the false negative rates, the ratio of actual positive people that were wrongly predicted as negative, to be equal between groups. Therefore, it is also called false negative error rate balance. When there false negative rates are equal between groups, then the true positive rates between groups are aslo equal. This means requiring equal false negative rates or equal true positive rates between groups results in the same effect. fulfilled, one could also define equal opportunity via the true positive rate. Predictive equality follows the same principle as equal opportunity but instead of focusing on the false negatives, it focuses on the false positives. Again, if a classifier has equal false positive rates between groups, it also has equal true negative rates. With its focus on the false positive rates, predictive equality is also presented in the context of punitive tasks. Since people could experience potential harm on the basis of a positive prediction, the proportion of truly innocent people that do not deserve punishment should be kept at a minimum. For assistive tasks, such as deciding who receives some kind of welfare, a focus on minimising the false negative rate could be more relevant. Equalized odds combines equal opportunity and predictive equality. It requires that the false positive and true positive rates are equal between groups, and is in this sense stricter than either of them alone. Treatment Equality is another variation that forms the error ratio for each group and requires it to be equal. Finally, overall accuracy equality simply requires equal accuracy between groups, meaning equal proportion of correctly classified individuals in each group. **Sufficiency** requires independence between Y and A conditioned on $\hat{Y}$, so $Y \perp A|\hat{Y}$. Intuitively this means that we want a prediction to be equally credible between groups. This leads to the following fairness metrics:

- Predictive parity/ outcome test: $P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b)$ `mlr3`: `fairness.ppv`
- Equal true negative rate: $P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = b)$ `mlr3`: `fairness.npv`

- Equal false omission rate: $P(Y = 1|\hat{Y} = 0, A = a) = P(Y = 1|\hat{Y} = 0, A = b)$ `mlr3: fairness.fomr`
- Equal false discovery rate: $P(Y = 0|\hat{Y} = 1, A = a) = P(Y = 0|\hat{Y} = 1, A = b)$
- Conditional use accuracy equality: $P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b) \wedge P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = b)$

Predictive parity requires that the probability of actually being positive, given a positive prediction is the same between groups. Following the same principle, we can require that the probability of actually being negative, given a negative prediction is the same between groups. If we instead look at errors again, we can require equal false omission rates between groups or equal false discovery rates between groups. False omission describes the case in which an actual positive person is predicted as negative and can be highly relevant in assistive settings, such as description of a medical treatment. False discovery rate describes the case in which an actual negative person is predicted as positive. This should be taken into account in punitive settings, in which we do not want to convict innocent people. By not only requiring one of these criteria but two simultaneously, we can build a stronger metric, like conditional use accuracy equality that requires same positive predictive values between groups and same negative predictive values between groups. Hopefully, the pattern becomes clear now. While it is easy to get overwhelmed by the amount of definitions at first, taking a closer look, it becomes clear that they are constructed in a structured way. In fact, equal false omission rate and equal false discovery rate were not introduced in the paper Verma and Rubin 2018 but it is clear that they follow the same pattern as the other metrics.
Most (binary) classifiers work with predictions scores and a hard label classifier is applied only afterwards in form of a treshold criterion. It should therefore come as no surprise that instead of formulating fairness with $\hat{Y}$ there exist fairness metrics that use the score S, which typically represents the probability of belonging to the positive class.

- Calibration: $P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b)$
- Well-calibration: $P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b) = s$
- Balance for positive class: $E(S \mid Y = 1, A = a) = E(S \mid Y = 1, A = b)$
- Balance for negative class: $E(S \mid Y = 0, A = a) = E(S \mid Y = 0, A = b)$

Calibration requires that the probability for actually being positive, given a score s is the same between groups. As the score can usually take values from the whole real number line, this can in practive be implmented by binning the scores Verma and Rubin 2018. Well-calibration is a stronger version of this, requiring that the probability for actually being positive, given a score s is the same between groups and equal to the score itself. This means, when for a set of suspects the classifier predicts a certain probability s of crime, then the proportion of people that actually committed crime should be s. Balance for the positive class takes the expectation over the predictions scores of the people that are actually positive and wants them to be equal across groups. We do not want that one the positive people of one group get on average a higher score than the positive people

of another group. The same holds for the negative class, formalized as balance for the negative class. To constrast the group fairness criteria, sufficiency takes the perspective of the decision-making instance, as usually only the prediction is known to them in the moment of decision. For example, the police, who do not yet know the true label at the time when they are supposed to decide whether someone would become a criminal. As separation criteria condition on the true label Y it is suitable when we can be sure that Y is free from any bias, so to say when Y was generated via an objectively true process (this will become clearer in the chapter on bias). Independence is best, when we want to enfore a form of equality between groups, regardless of context or any potential personal merit. While this seems be useful in cases in which the data contains complex bias, it is unclear whether this enforcments has the intended benefits, especially over the long term. reference?

Individual fairness

If we want to equalise e.g. the false positive rates between two groups and currently group a has a higher false positive rate than group b, this would lead us to lowering the prediction threshold for b, such that more actual negative people would get classified as positive. Or it we would need to set a higher thresholkd for group a, such that it becomes harder for them to be classified as positive. Depending on the context, either option can seem unfair. Individual metrics therefore shift the focus. The underlying idea of fairness is that similar individuals should be treated in a similar way. Different individuals should be treated in a different way. It is an intuitive idea that was already formulate by greek philosopher bothmann citation. **Fairness through awareness (FTA)** formalizes this idea as Lipschitz criterion.

$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$

This simply puts an upper bound to the distance between predictions of two individuals, which depends on the features of them. In other words, if two people are similar in their features, they should also get similar predictions from the algorithm. The challenge of FTA is the definition of precisely this equality in the feature space. Defining when two individuals are similar is not much different from defining fairness in the first place Castelnovo et al. 2022. Threre is no clear solution to this. In any case, the choice of d_X should take context-specific information into account. We want to find a distance metric, that suits the target and represents an ethical formalisation of similarity in the features.
Fairness through unawareness (FTU) or blinding This is primarly formulated as procedural rule. Blinding tells us to not use the protected attribute explicitly in the decision-making process. So at first this would simply mean to discard the protected attribute from the data during training. After training, FTU can be tested by simulating a doppelganger for each person in the dataset. This doppelganger has the exact same features with the excpetion of the protected attribute, which is flipped (easy in binary PA case). If both these instances have the same prediction, the algorithm would satisfie FTU Verma and Rubin 2018. This is actually also a from of FTA, in which we chose the distance metric to measure a distance of zero only if two people are the same on all their features except for the protected attribute. Blinding is however to be seen critically. FTU has the problem of proxies. These are variables that are strongly correlated with

the protected attribute. Therefore, it's not enough to simply mask the information of the sensitive attribute druing training because discrimination can persist via these proxis. This becomes clearer, when imagine that we remove information, such that this feature is simply not available to the classifier during training. The place of residence, however, is strongly correlated with the person's ethnicity. Thus, indirect discrimination based on ethnicity remains, even though the information was not directly available during training. Suppression therefore extends the idea of blinding and the goal is to develope a model that is blind to the sensitive attribute and the proxis. The drawback is, that it is unclear when a feature is sufficiently correlated with the sensitive attribute to be counted as proxi. Additionally, we could loose important information by removing too many these features Castelnovo et al. 2022.

Causality-based notions

In contrast to observational fairness metrics, causality-based notions ask whether the sensitive attribute was the *reason* for the decision. If a certain (harmful) decision was made *because of* the value of the sensitive attribute of a person, we deem the algorithm as unfair. There are causality-based concepts that focus on group-level fairness and also some that focus on individual-level fairness. We want to give an intoduction to all of them, but since this category requires a new theory we will not get into great detail.
Group-level: FACE, FACT (on average or on conditional average level) (Zafar et al. 2017)
Individual-level: counterfactual fairness, path-based fairness (Kusner et al. n.d.) The two most common individual fairness metrics are counterfactual fairness and path-based fairness.

Stop, Question, and Frisk data

The legal sector is one in which ADMs have been deployed, more often than not accompanied by public debate and protests about targeted policing and racial discrimination (COMPASS as the most popular example). We will turn our focus to the stop, question, and frisk (SQF) dataset published by the New York police department (NYPD). A. Fabris and S. Messina and G. Silvello and G.A. Susto recommend it as suitable dataset for fairness research. First, we will give some context to the dataset. We will continue with descriptive analysis and finally examine fairness of an algorithm trained on this data.

Background

Since x the stop, question, and frisk practice is implemented in New York City. A police officer is allowed to stop a person if they have reasonable suspicion that the person has committed, is committing, or is about to commit a crime. During the stop the officer is allowed to frisk a person (pat-down the person's outer clothing) or search them more carefully. The stop can result in a summon, an arrest or no further consequences. After a stop was made, the officer is required to fill out a form, documenting the stop. This data is published yearly by the NYPD. Many citizens have critisised the stop and frisk practice. There is disagreement about whether the strategy is effective in reducing the crime rates

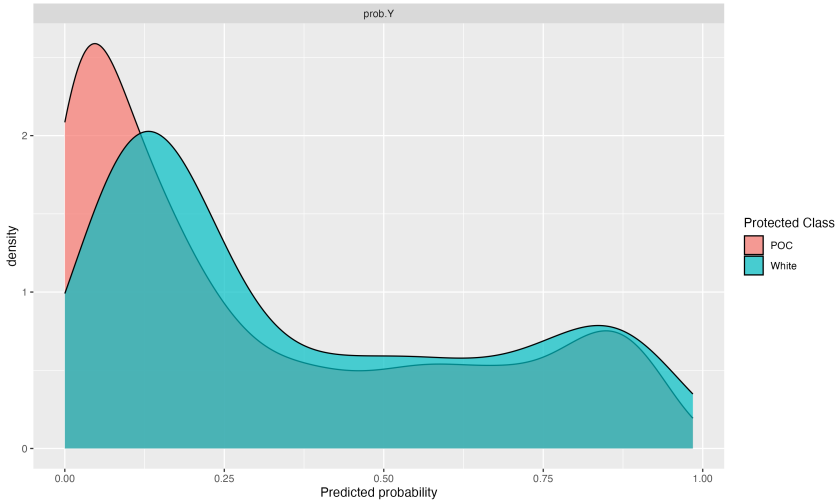

Figure 1: Density of predicted probabilities both groups.

of the city. The police has been repeatedly critisised for over-targetting people of colour. Stop and Frisk practice during 2004 to 2012 has been deemed as unconstitutional. source

Data description

For our analysis we look at the stops from 2023, which were the most recent recordings to the time of writing this paper. The raw 2023 dataset consists of 16971 observations and 82 variables. We first discarded all the variables that have more than 20% missing values. 34 variables remain that provide us with information about the stop and demographic information of the stopped person. From this reduced dataset we filter out the complete cases and end up with 12039 observations. [1] We choose the arrestment of a suspect as target and the race as protected attribute. To adjust our situation to the common binary classification, binary PA scenario in the fairness literature, we dichotomise the race attribute by grouping "Black", "Black Hispanic", and "White Hispanic" as people of colour ("PoC") and "White", "Asian Pacific", and "American Indian/Native American" as white ("White").

Fairness Auditing

We train a regular random forest classifier to model the decision to arrest a person. With many of the group fairness metrics implemented in mlr3fairness, we can measure the (group) fairness of our models. We find that the random forest classifier is already fair. There are minor differences between groups, but exact equality cannot be expected in practice, thus it is common to allow for a certain margin of error ϵ. Especially the error rates (fnr, fpr) are very similar between groups, thus Separation seems to be satisfied overall. Sufficiency metrics have larger differences, though they are still minor. From

[1] Simply discarding the missing values and only training on complete cases is discouraged by Fernando et al. 2021. We opt for this approach regardless, since imputation of the missing values is not straight forward but treating missing values as an extra category (which some random forest learners in mlr3 can do) will introduce complications when we implement some fairness methods later on.

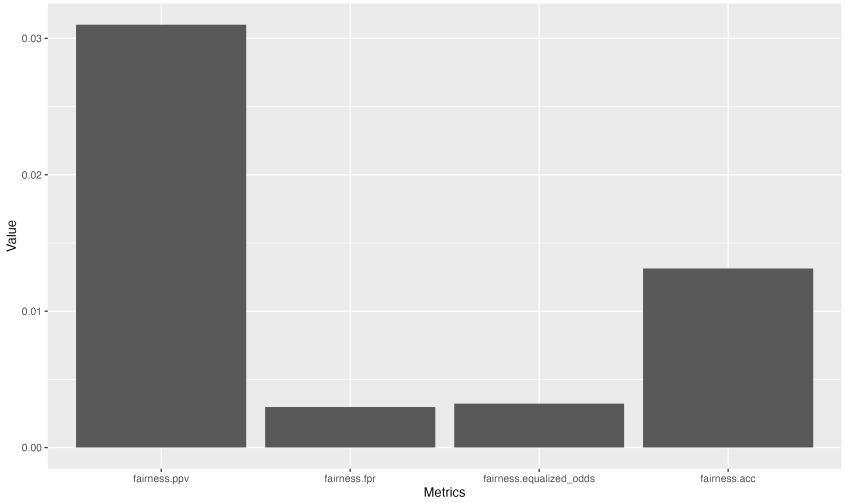

Figure 2: Comparison of fairness metrics.

Figure 2 we can directly see that the positive predictive value has a relatively large difference between groups, while the false positive rate is practically the same between groups (Separation). It comes to no surprise that equalised odds, which is based on error rates, is satisfied. Finally, the accuracy between groups is not as equal as the error rates are, but the absolute difference is still smaller than 0.05, which is a common ϵ to choose. Given that the classifier is fair from a group perspective, it does not make sense to experiment with any of the implemented fairness methods in mlr3. At most, we could try to address the disparities in sufficiency metrics, but the common methods in the package are designed to address concerns with independence or separation.
What is more interesting, is to compare the classifier trained on data from the unconstitutional period 2004 to 2012. We decide for 2011 as it is the year with the most stops. We carry out the same data cleaning steps for the 2011 data as before, starting with 685724 recorded stops and reducing this to 651567 clean observations. Note, these are more than 50 times more stops than in 2023. While this means that the 2011 data has substantially more low-risk stops the racial disparities are interestingly even smaller than for the 2021 data. Our fairness audit did not show any substantial disparities in fairness metrics. Does this mean the classifier is fair? It is easy to come to such conclusions, especially if fairness is not the major concern of the practitioners but more of a nuisance criterion that should be fulfilled. However, to truly ensure a fair practice, it is crucial to look at the context in which the algorithm is embedded.
Usually fairness is a concern in the first place, because the algorithm should be implemented as an ADM to assist decision-making in some way. As such it could influence if someone gets admitted to college, gets a loan or is released from prison. The algorithm therefore does not exist in isolation, but is embedded in a loop with data and the user. We make the circumstances of a decision measurable by collecting data. The algorithm learns from this data to make an optimal prediction, on which the decision-makers base their judgement on Figure 3. At each step of this loop, bias can be introduced in the process and even be amplified as the algorithm influences decision-making on a large scale. This means that every fairness project comes with the task to understand where the data

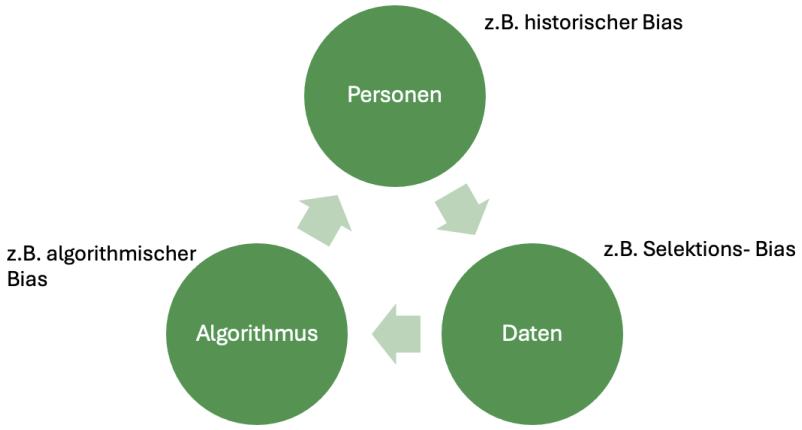

Figure 3: The bias loop.

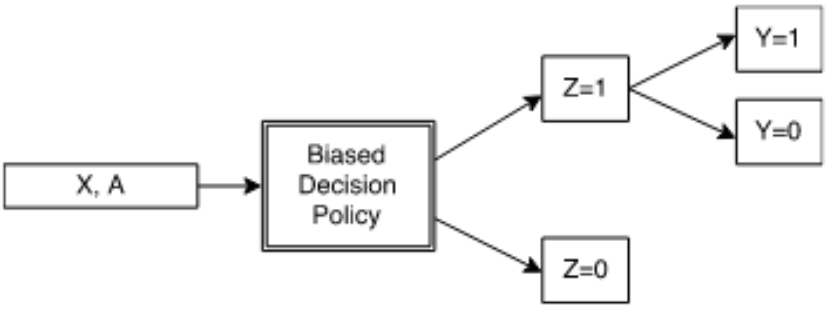

Figure 4: Selection bias in the SQF data.

comes from and how exactly the algorithm will be deployed in practice. Let us therefore take a step back and look at the context of the SQF data.

Sources of bias in the SQF data

The major concern that has been identified in the literature for SQF data is how the data is generated in the first place. In their paper "Residual Unfairness" Kallus and Zhou n.d. conceptualise the problem as shown in Figure 4. We define a person by their sensitive feature (A) and non-sensitive features (X). For each person in the population of interest a police officer decides whether to stop them or not. Based on this biased decision policy people are included in the sample (Z = 1) or they are not (Z = 0). But naturally we only observe further information on the people that were stopped. Only for them we can know the outcome of a stop, which constitutes the target of a classificaiton task. Kallus and Zhou n.d. distinguish between target population and training population in such scenarios. The target population is the one on which we want to use the ADM on while the training population are the observations the biased decision policy chose to include in the sample and on which the algorithm is trained. In the SQF data we can see this form of bias by comparing the race distribution of NYC to the race distribution in the SQF

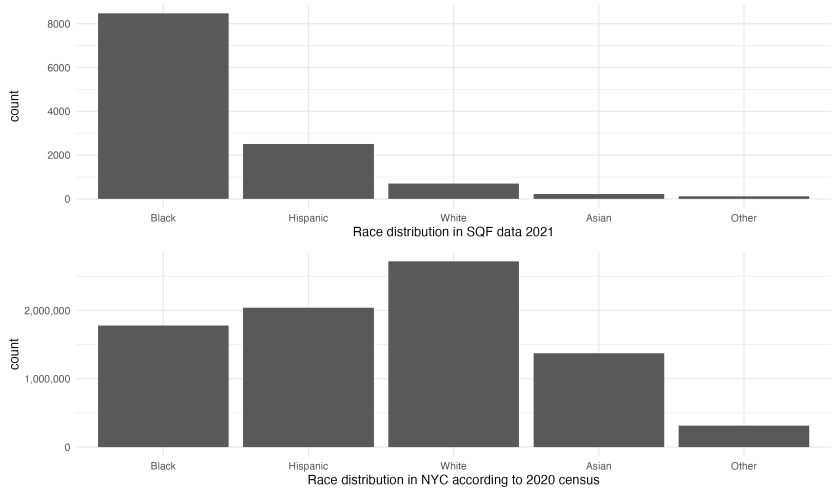

Figure 5: Comparison of race distribution in the training and target population.

data. From Figure 5 it is clear that in terms of race the SQF data does not represent the general population of the city. We can see that white people form the majority of the population in NYC, but only make up a tiny fraction of SQF stops. Black people in contrast are the third-largest ethnic group in NYC while they exceed any other group in the SQF data by far. Figure 5 shows that selection bias might be at play in the decision of stopping a suspect.
To get a more nuanced picture we also plot the true arrestment rate per ethnic group in the SQF 2023 data and find that White and Asian people have the highest arrestment rates and black people the lowest. This supports the argument that black people are stopped more leniently and there is a biased decision policy in place. The questions now are why the group metrics did not detect any unfairness in our algorithm and how such biases can be addressed in fairness practice. Why the group metrics did not show any unfairness in the algorithm can be answered ins a straight forward way. Group metrics offer a rather isolated view on fairness. They assess disparities in algorithimic predictions between protected groups rather than measuring the fairness of a whole situation. Thus, group metrics are not designed to detect selection bias. They work with the joint distribution of $Y, A, X, \hat{Y}$ and do not take any additional information into account. When we rely on the true label Y (Separation) to detect unfairness but the true label itself is not reliabel (generated via an objective truth), then the group metrics cannot show this Castelnovo et al. 2022. To answer the second question we use the following chapter to examine two papers that propose methods to address selection bias in fairness practice and have explicity used the SQF data as a case study.

2 Residual Unfairness

We already outlined the formal problem setting in Kallus and Zhou n.d. in Figure 4. The main message of the paper supports the saying "bias in, bias out". They argue that fairness adjustments on the training population do not ensure fairness in the target

population. Fairness is defined via equal opportunity or equalised odds. They propose to modify a post-processing method to deal with selection bias.
The paper uses the following notation: A is the binary sensitive attribute, X are the non-sensitive attributes, Y is the target, $\hat{Y}$ is the prediction, Z is the decision of the biased inclusion policy ($Z = 1$ means the subject is included into the training population). T indicated whether a person is in the target population ($T = 1$) means that we want to use the trained algorithm on the person. $\hat{R} \in [0, 1]$ is the prediction score. The problem depicted in Figure 4 can be now expressed as follows. We only have knowledge about $X, A, Y|Z = 1$ while we do not know $X, A, Y|Z = 0$.
The paper defines fairness via equal opportunity or equalised odds. Recall that equal opportunity demands that the false positive rate is the same across groups, i.e. $P(\hat{Y} = 0|Y = 1, A = a) = P(\hat{Y} = 0|Y = 1, A = b)$. But requiring equal false negative rates across groups is the same a requiring equal true positive rates across groups since $P(\hat{Y} = 0|Y = 1, A = a) = 1 - P(\hat{Y} = 1|Y = 1, A = a)$. The paper uses a thresholding classifier, if the prediction score exceeds a certain threshold the positive value for the target is predicted. This allows us express the false negative rate and the true positive rate of a group a with respect to an event E via the cummulative distibution function. $F_a^E = P(\hat{R} \leq \theta|Y = 1, A = a, E)$. Note that we condition on the truly positive subject in the sample. In the SQF case this would mean that we only look at people that were truly arrested. The truly arrested that have a predicted probability $\hat{R} \leq \theta$ are wrongly classified as not-arrested, while the ones for whom $\hat{R} > \theta$ are correctly classified as arrested. F_a^Z gives us nothing other than the false negative rate in the training population and F_a^T is the false negative rate in the target population. So when we want to define an equal opportunity classifier on the training population we require $F_a^Z(\theta_a) = F_b^Z(\theta_b)$ to hold.
An optimal derived equal opportunity classifier can then be defined as $\hat{Y} = I(\hat{R} > \theta_A)$ and $F_a^Z(\theta_a) = F_b^Z(\theta_b)$ for all groups a,b. In words, the classifier predicts the positive outcome with a group-specific threshold for each member of the group while this group specific threshold is set in such a way that equal opportunity on the training data is fulfilled. We will not go into detail of how one finds such a classifier but refer to Hardt et al. 2016 who propose a post-processing method to derive the optimal thresholds for each group.
With the definition of fairness as equal opportunity (equal tpr/fnr across groups) we can in turn define unfairness as inequity of opportunity. This describes nothing other than the difference in true positive rates between groups, i.e. $\epsilon_{a,b}^E = P(\hat{Y} = 1|Y = 1, A = a, E) - P(\hat{Y} = 1|Y = 1, A = b, E)$. $\epsilon_{a,b}^{T=1} > 0$ shows discrimination against group b, since this means that the true positive rate for group b is lower than for group a. When we construct an equal opportunity classifier (via some fairness intervention) then $\epsilon_{a,b}^{Z=1} = 0$ holds for this classifier. This also means that any inequity of opportunity that might show in the target population cannot be explained via existing ineuqities between groups in the training population but via existing differences in the training and the target population. So it is unfairness that gets introduced when we try to generalize our algorithm to the population it was not trained on. Kallus and Zhou n.d. call this residual unfairness, since this is unfairness remaining event after fairness adjustments.

Strong disparate benefit of the doubt

To fully understand the results of the paper, we additionally introduce the concept of stochastic dominance, originating from decision theory. Let F, G be two cummulative distribution functions. Then G first order stochastically dominates $F \preceq G$ when $F(\theta) \geq G(\theta)\ \forall\theta$. Recall that G and F are cummulative distribution functions. So first order stochastic dominance of G over F, smaller values of G for each input value θ, that the population described by the CDF of G consistently has higher probability values than population F. Their probability mass is concentrated towards the higher input values thus the cummulative distribution function is small for small input values. Equipped with these definitions the paper constructs difference scenarios of (un)fairness. The bottom line always is that equal opportunity in the training population does not guarantee equal opportunity in the target population.
For the first discrimination scenario we assume the following: $F_a^{Z=1} \preceq F_a^{T=1} and F_a^{Z=1} \succeq F_a^{T=1}$ and at least one of the equalities does not hold (either $F_a^{Z=1} \neq F_a^{T=1}$ or $F_b^{Z=1} \neq F_b^{T=1}$ or both). In words this means for group a, the target population first order stochastically dominates the training population. The truly positive individuals in group a consistently have lower scores than the ones in the target population. The score is the predicted probability of receiving the positive outcome. When for group a we have more people with low probabilites for the favourable outcome in our sample, then this means that group a member were more leniently introduced into the sample. The paper says that group a members recevied more "benefit of the doubt". For the SQF scenario we reverse things. $\hat{Y} = 1$ is no longer desirable and we can interpret scores as riskscores G_g^E. So the score describes the probability for receiving the undesirable outcome. The proposition can be reformulated by switching the order signs. Suppose that $G_a^{Z=1} \preceq G_a^{T=1}$ and $G_b^{Z=1} \preceq G_b^{T=1}$ while at least one of the equalities does not hold i.e. $G_a^{Z=1} \neq G_a^{T=1}$ or $G_b^{Z=1} \neq G_b^{T=1}$ or both. Then every derived equal opportunity classifier has nonnegative inequity of opportunity for group b relative to group a $\epsilon_{a,b}^{T=1} \geq 0$ and at least one derived equal opportunity classifier will have a strictly positive inequity of opportunity disadvantaging group b relative to group a $\epsilon_{a,b}^{T=1} > 0$. In the context of SQF this means that group a members were stopped more carefully while group b members were stopped leniently. This aligns with the fact that the sqf data records considerably more stops for black people than white people. In this case the propositions of the paper will show us again that even after adjusting for equal error rates, the classifier will disadvantage group b when applied to the target population.
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 group 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 interpret 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 $G_b^{Z=1} \preceq G_b^{T=1}$. When we apply the algorithm to the target population 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 unsure if this makes sense, if a transfered it correctly.
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?
Chapter 6: Case study on SQF data Their main message is always, bias in, bias out. fairness interventions, done on the training 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 [2] 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.

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.
My big questions is is these two papers are actually contradicting each other. I think they do not. What both are essentially saying is that if the distribution of the target in training and target population is different, then there will be a problem. Kallus and Zhou n.d. looks at the situation in which training and target datahave different distributions in both groups. In the stricter scenario the difference in target and train exists for both gorups and is going in opposite directions (e.g. train of a is underestimation and train of b is overestimation). In the relaxed scenario the difference in target and train exists for both groups but goes in the same direction, I think is is just more severe for the disadvantaged group. In Rambachan and Roth n.d. we say that our limited sample is not biased necesariliy in itself in the sense that the distribution of the target in our sample is different from the distribution of the target in the target population per se. But what happens is that the sample is limited and therefore only cuts out a piece of the target population that is not representative. Therefore, when we collect more data we come

[2] 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.

closer to the target population and our classifier will work better on the target population for the group with more data.
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

Castelnovo, Alessandro et al. (Mar. 2022). "A Clarification of the Nuances in the Fairness Metrics Landscape". In: *Scientific Reports* 12.1, p. 4209. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07939-1. (Visited on 12/23/2024).

Caton, Simon and Christian Haas (July 2024). "Fairness in Machine Learning: A Survey". In: *ACM Computing Surveys* 56.7, pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3616865. (Visited on 12/23/2024).

Corbett-Davies, Sam et al. (n.d.). "The Measure and Mismeasure of Fairness". In: ().

Fernando, Martínez-Plumed et al. (2021). "Missing the Missing Values: The Ugly Duckling of Fairness in Machine Learning". In: *International Journal of Intelligent Systems* 36.7, pp. 3217–3258. ISSN: 1098-111X. DOI: 10.1002/int.22415. (Visited on 12/10/2024).

Hardt, Moritz et al. (2016). "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc. (Visited on 01/27/2025).

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

Kusner, Matt J et al. (n.d.). "Counterfactual Fairness". 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).

Zafar, Muhammad Bilal et al. (2017). "From Parity to Preference-based Notions of Fairness in Classification". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. (Visited on 12/29/2024).

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, February, 12th 2025

Juliet Fleischer