Seminar Thesis

FairML and the SQF dataset

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, February, 25th 2025

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

Abstract

In the first half of this paper we provide an introduction to the most common metrics and methods in fair machine learning. We then apply the theoretical concepts to the New York Stop, Question and Frisk dataset, which will showcase difficulties that come with fairness in practice. This leads us to explore the problem of selection bias and related issues. We turn our focus to studies that have worked with the SQF dataset and established interesting theoretical results; residual unfairness, bias reversal and bias inheritance. Value of this paper: compare and contrast traditional fairness metrics, use them in a real world setting, show its limitations in this setting, present how they are adressed in more advanced ways and try to explain why traditional metrics can not reflect the whole situation.

Contents

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A \vert Y$	$Y \perp A \vert \hat{Y}$

1 Introduction

2 Related Work

Badr and Sharma 2022
Rambachan and Roth 2020
Kallus and Zhou 2018
Goel, Rao, and Shroff 2016
Khademi et al. 2019

3 Fairness Metrics

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.

1. Group fairness vs. individual fairness
2. observational vs. causality-based criteria

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 random variables characterizing the population of interest to assess fairness while causality-based criteria make assumptions about the causal structure of the data and base their notion of fairness on this. On the basis of these fundamental ideas, a plethora of formalizations 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 extension 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. To bring more clarity we will illustrate the general fairness metrics already in the context of the SQF dataset.

3.1 Group fairness

The notion of fairness underlying group metrics is that discrimination of certain groups of the population defined via the PA should be prevented. The groups metrics presented here are observational metrics. They can be separated into three main categories, independence, separation, and sufficiency.

Independence

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 $\hat{Y} \perp A$. This is fulfilled when for each group the same proportion is classified as positive by the algorithm. 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 formalized as
demographic parity/statistical parity

$$P(\hat{Y}|A = a) = P(\hat{Y}|A = b)$$

This indeed seems like a simple, perhaps too simple definition of fairness, as it only looks the distribution of $\hat{Y}$ conditioned on the group membership. There are extensions of this idea, such as conditional statistical parity. This metric allows to condition on A and a set of legitimate features E. For instance, predictive parity would mean that we require equal prediction ratios between PoC and white people while conditional statistical parity requires equal prediction ratios between PoC and white people who *live within the same borough of New York* (E = borough). This can be seen as a more nuanced approach, as it allows tacking additional information into account. The other two groups of group fairness metrics, Separation and Sufficiency can both be derived from the error matrix.

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	TN	FN
$\hat{Y} = 1$	FP	TP

Separation

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 $P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$

- Predictive equality/ False positive error rate balance

  $$P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b)$$

  or
  $P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$

- Equalized odds

  $$P(\hat{Y} = 1|Y = y, A = a) = P(\hat{Y} = 1|Y = y, A = b) \forall y \in \{0, 1\}$$

- Overall accuracy equality:

$$P(\hat{Y} = Y|A = a) = P(\hat{Y} = Y|A = b)$$

- 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 also equal. This means requiring equal false negative rates or equal true positive rates between groups results in the same effect. 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. 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.
In itself, these error rates are detached from the context of fairness and used in general in machine learning to assess the performance of a classifier. In essence the group metrics we outlined so far do nothing other than picking a performance metrics from the confusion matrix and requiring it to be equal between two (or more) groups in the population. This means the well-known trade-offs for example between false positive and true positive rate are also present in the fairness metrics. As more people get correctly classified as positive usually also more people get wrongly classified as positive. Source With this comes the difficulty to choose "the right" metric for the specific task. In general one can think about this in the same way as when choosing a performance metric for a binary classifier. In setting in which a positive prediction leads to a harmful outcome, as in the SQF setting, it often makes sense to focus on minimizing the false positive rate, while a higher false negative rate is accepted as a trade-off. This argumentation follows the idea of. The authors distinguish between punitive and assistive tasks to help choose the right fairness metric. For punitive tasks metrics that focus on false positives, such as predictive equality are more relevant. For assistive tasks, such as deciding who receives some kind of welfare, a focus on minimizing the false negative rate could be more relevant, so equal opportunity would be more suitable.

Sufficiency

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. When a white person gets a positive prediction the probability that it is correct should be they same as for a black person. This leads to the following fairness metrics:

- Predictive parity/ outcome test requires that the probability of actually being positive, given a positive prediction is the same between groups.

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b)$$

- Equal true negative rate follows the same principle as predictive parity. It requires that the probability of actually being negative, given a negative prediction is the same between groups.:

$$P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = b)$$

- If we instead look at errors again, we can require equal false omission rates:

$$P(Y = 1|\hat{Y} = 0, A = a) = P(Y = 1|\hat{Y} = 0, A = b)$$

- Or equal false discovery rate:

$$P(Y = 0|\hat{Y} = 1, A = a) = P(Y = 0|\hat{Y} = 1, A = b)$$

- Conditional use accuracy equality:

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = b) \wedge P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A =$$

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. Just as for equalized odds as Separation metric, we can build a stronger Sufficiency metric requiring multiple conditions to hold simultaneously, e.g. conditional use accuracy equality. 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 are implemented in `mlr3fairness`, and 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 threshold 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. Instead of conditioning on $\hat{Y}$ as Separation metrics, we can simply condition on S and define:
Calibration:

$$P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b)$$

Calibration requires that the probability for actually being positive, given a score s is the same between groups. So the idea is a more fine-grained version of predictive parity. As the score can usually take values from the whole real number line, this can in practice be implemented by binning the scores Verma and Rubin 2018.
To compare 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 enforce a form of equality between groups, regardless of context or any potential personal merit. While this seems to be useful in cases in which the data contains complex bias, it is unclear whether these enforcements have the intended benefits, especially over the long term. Reference?. It is good to understand the difference in perspectives each of the group fairness metrics take, because many of them cannot be satisfied simultaneously. This is known as the impossibility theorem Hardt et al. 2016. This means one has to decide on either Independence, Separation or Sufficiency and the choice should fit the context of the data and the decision-making process. Lastly, we note that these are not all the group fairness metrics that exist, but broadly speaking other metrics are variations of the presented ones. Some more metrics are listed in the appendix.

3.2 Individual fairness

If we want to equalize 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 if we would need to set a higher threshold for group a, such that it becomes harder for them to be classified as positive. Depending on the context, either option can seem unfair. So by trying to equalize a given metric between groups, it can happen that individuals within a group are treated unequally. Individual metrics therefore shift the focus. The underlying idea of fairness is that similar individuals should be treated similarly.

Fairness through awareness (FTA)

FTA formalizes this idea as Lipschitz criterion.

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

d_Y is a distance metric in the prediction space, d_X is a distance metric in the feature space and λ is a constant. The criterion 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 close in the feature space, they also should be close in the prediction space. The challenge of FTA is the definition of the equality in the feature space. Castelnovo et al. 2022 states "defining when two individuals are similar is not much different from defining fairness in the first place". In the SQF context, it could make sense to define similar individuals based on yearly income, age and neighbourhood. Yet one could easily argue that taking the criminal history into account is important as well. After the decision for a legitimate set of features has been made, the next challenge is to choose a distance metric that appropriately captures the conceptual definition of similarity defined via the selected features. FTA does not have one clear solution and requires domain knowledge and the choice of d_X should take context-specific information into account.

Fairness through unawareness (FTU) or blinding

In contrast to FTA, blinding should give a simple, context-independent rule. It tells us to not use the protected attribute explicitly in the decision-making process. When training a classifier this means discarding the PA during training. Since FTU is a more procedural rule than a mathematical definition, there exist multiple ways to test whether the blinding worked for a classifier. One approach is to simulate a doppelgänger for each observation in the dataset. This doppelgänger has the exact same features except the protected attribute, which is flipped. If both these instances have the same prediction, the algorithm would satisfy FTU Verma and Rubin 2018. [1] Other ways to assess FTU can be found in Verma and Rubin 2018. A problem blinding has been proxies. These are variables that are strongly correlated with the sensitive attribute. It is not enough to simply mask the information of the sensitive attribute during training because discrimination can persist via these proxies. For SQF this would mean that we remove the race attribute during training. A person's ethnicity, however, is strongly correlated with their place of residence. Thus, indirect discrimination based on ethnicity remains, even though the information was not directly available during training. **Suppression** extends the idea of blinding and the goal is to develop a model that is blind to not only the sensitive attribute but also the proxies. The drawback is, that it is unclear when a feature is sufficiently high correlated with the sensitive attribute to be counted as proxy. Additionally, we could lose important information by removing too many features Castelnovo et al. 2022.

3.3 Causality-based fairness metrics

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.
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. 2017) The two most common individual fairness metrics are counterfactual fairness and path-based fairness.

3.4 Comparison and Summary

The difference between observational and causal clear, really different approach. The division in group and individual fairness metric actually more of a nuanced differentiation. The observational metrics can rather be ordered on a plane, depending on how much information of the situation via other features X they allow. Traditional group metrics like demographic parity, equal error rate metrics and sufficiency metrics only work with the distribution of $Y, \hat{Y}, X, A$. The individual fairness metrics take more information of the non-sensitive feature into account in order to define similarity. Metrics such as

[1] This can be seen as 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. In this special case FTA and FTU are measured in the same way.

conditional demographic parity lie in between, as we allow for a relevant subset of non. Sensitive feature to be part of the definition. Castelnovo et al. 2022 therefore depict this as a plane. The amount of approaches to measure fairness shows the complexity of the topic. There is not *the* right fairness metric to choose, but there can be the best one depending on the context and the data. The next section will present ways to digitate algorithmic bias once detected by one of the fairness metrics.

4 Fairness Methods

Fairness methods

Another question fair machine learning deals with is how algorithms can be adjusted such that they fullfil one of the above fairness metrics. Depending on when they take place in the machine learning pipeline, we distinguish between preprocessing, inprocessing or postprocessing methods. Preprocessing methods have the idea that the data should be modified before training, so that the algorithm learns on "corrected" data. Reweighing observations before training is an example for a preprocessing method, that we will use in our case study in chapter x. In Processing methods modify the optimisation criterion, such that a it also accounts for a chosen fairness metric. Introducing a regularization term to the loss function is one example of such modifications. Postprocessing methods work with black box algorithms, just like preprocessing methods. We only need the predictions from the model to adjust them so that again a chosen fairness metric is fullfilled. One example for this is thresholding, where we set group specific thresholds to re-classify the data after training. We will discuss a post-processing approach by Hardt et al. 2016 in a following chapter.

5 Case Study: Stop, Question, and Frisk

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 scanned more than x datasets to diversify the datasets that are used in the fairness literature. They 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.
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 criticized the stop and frisk practice. There is disagreement about whether the strategy is effective in reducing the crime rates of the city cite some studies. The police has been repeatedly criticized 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 as they were the most recent at 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, which leaves 34 variables. From this reduced dataset we filter out the complete cases and end up with 12039 observations. [2]
Race is the protected attribute (PA). For the fairness audit later in the chapter we dichotomize the PA to adjust our situation to the common binary classification, binary PA scenario in the fairness literature. For a more nuanced descriptive analysis we only summarize "Black Hispanic" and "Black" into the group "Black" and "American Indian/ Native American" and "Middle Eastern/ Southwest Asian" into the "Other" category. Black people are by far most often stopped, making up 70% of the total stops; yet, according to 2020 census data black people make up only 20% of the city's population Figure 1. At the same time white people form the majority of New York citizens (30%) but contribute with only 6% to the stops. After 2021 there has been a stark decline in stops and the police is known to focus their attention on high crime areas. Therefore, we further look at each borough. The most stops in 2023 occur in Bronx and Brooklyn. Based on report of the NYPD and population statistics from 2020, the Bronx also has the highest estimated crime rate per 100,000 citizens. Manhattan is not far behind in crime rate, but has fewer stops. Note that Bronx and Brooklyn happen to be the boroughs with the highest proportion of black citizens Figure 3.
After a more general overview of the dataset, we turn to the outcome of the stop. Specifically, we are interested in the arrestment of a suspect. In the cleaned 2023 data about 31% of stops result in an arrest. Overall racial disparities in arrestment rates are low. The arrestment rate for white suspects is the highest. **??**. As group fairness metrics are observational and constructed from the joint probability of $Y, \hat{Y}, A$, this already gives us a hint that the classifier trained to predict the arrestment of a suspect might show little racial disparities.

Fairness Auditing

To train a random forest classifier on the 2023 data to predict the arrest of a person, we dichotomize the race attribute by grouping "Black" and "Hispanic" as people of colour ("PoC") and "White", "Asian", and "Other" as white ("White"). As features, we select variables that should resemble the information that were available to the officer at the time

[2]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.

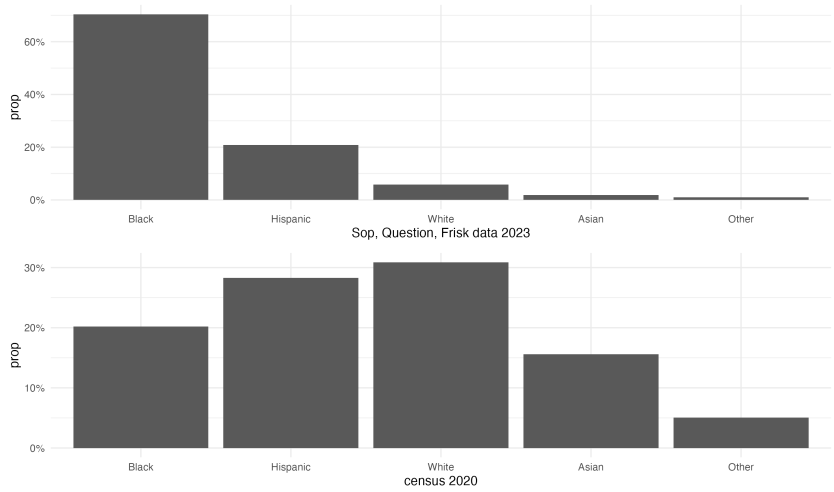

Figure 1: Comparison of race distribution in the training and target population.

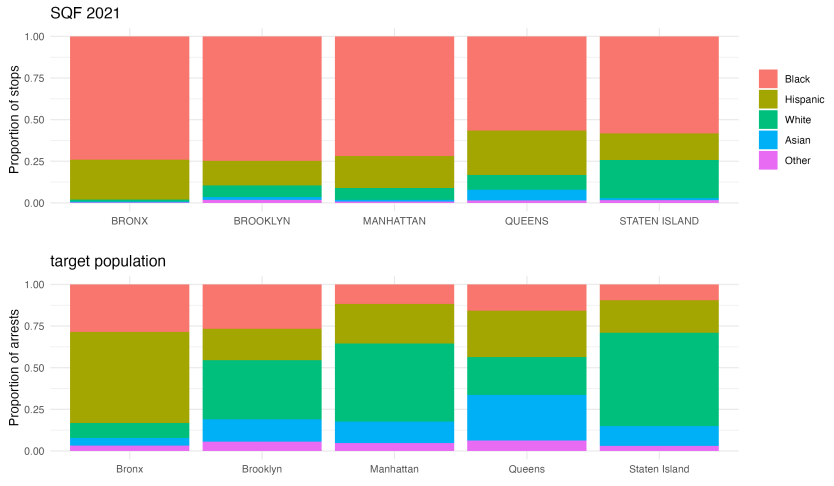

Figure 2: Racial distribution in the SQF data of each borough in comparison to the racial distribution of each borough in NYC as a whole.

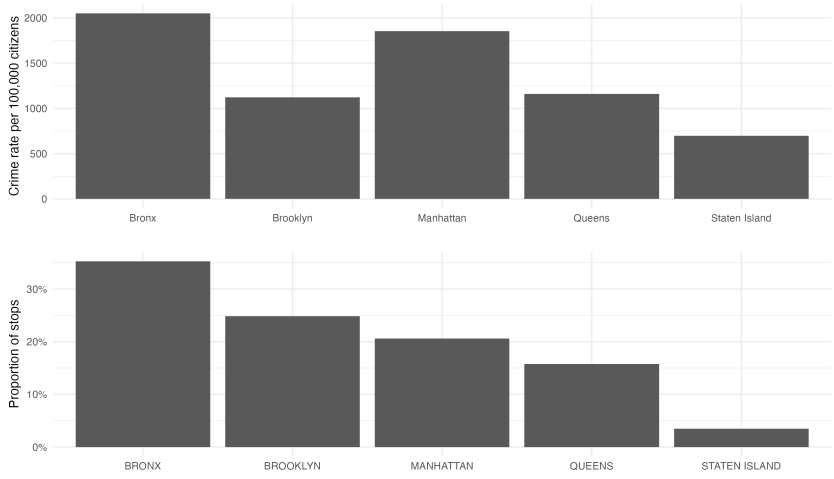

Figure 3: Racial distribution in the SQF data of each borough in comparison to the racial distribution of each borough in NYC as a whole.

of the stop. Additionally, we control for factors, such as the time of the stop or whether the officer was wearing a uniform. This selection of features is inspired by previous studies on SQF data (e.g. Goel, Rao, and Shroff 2016). With many of the group fairness metrics implemented in `mlr3fairness`, we can measure the (group) fairness of the regular random forest classifier.

First we plot the prediction score densities for each group in Figure 4. We can see that in general white people tend to have higher predicted probabilities than PoC. There is a relatively large proportion of PoC with low predicted probabilities, as seen by the peak at around 0.05. Recall that the predicted probability resembles the probability of being predicted positive (arrested). In Figure 5 we plot the absolute difference in common group fairness metrics. Exact equality of the group metrics cannot be expected in practice, so it is common to allow for a margin of error ϵ. Taking $\epsilon = 0.05$, the classifier is fair according to equalized odds, predictive parity, and equal opportunity. The classifier is unfair according to overall accuracy equality and predictive equality. This again shows that there is not one clear solution to fairness, but it depends on the perspective one takes.

With the absolute differences alone, it is possible to quantify the magnitude of unfairness but not the direction of it. Thus, we report the group-wise error metrics in Table 1. White people have a higher true positive rate, but also a higher false positive rate. This is usually a natural trade-off in classification tasks. For PoC the negative predictive values and the positive predictive value is higher. This means that given a person of colour is predicted as arrested, it is more likely that they were indeed arrested. The same for a negative prediction. Additionally, the false discovery rate, the proportion of predicted positives that are actually negative, is lower for PoC. This means that the classifier is more conservative in predicting positive outcomes for PoC. Overall, the accuracy is higher for PoC. Apart from the lower true positive rate, one could say that the classifier performs better for PoC than for white people. We will leave this interpretation as it is for now and will return to an alternative view later on.

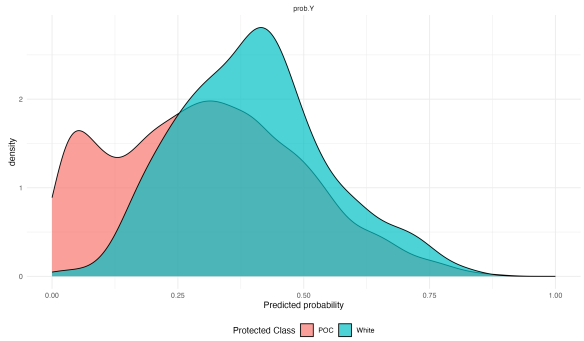

Figure 4: Density of predicted probabilities for both groups.

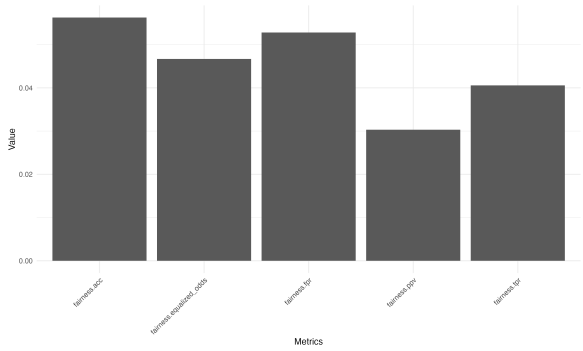

Figure 5: Another relevant plot.

Metric	PoC	White
tpr	0.34	0.38
npv	0.75	0.69
fpr	0.07	0.13
ppv	0.68	0.65
fdr	0.32	0.35
acc	0.74	0.68

Table 1: Groupwise Fairness Metrics (2023)

Fairness Experiment

Given that there are indeed disparities for two of the chosen metrics that exceed the ϵ threshold, we experiment with some fairness methods for the SQF data. The `mlr3fairness` package currently has two pre-processing methods, one post-processing method and some fairness adjusted models implemented. We decide to use a reweighing methods that works with assigning weights to the observations to equalize the distribution of $P(Y|A)$. The in-processing method is a fairness-adjusted logistic regression implemented in `mlr3fairness` inspired by Zafar et al. This method optimizes for statistical parity (Independence). The post-processing method aims for equalized odds, and it works by randomly flipping a subset of predictions with pre-computed probabilities in order to satisfy equalized odds constraints. reference to mlr3fairness book
In Figure 6 we compare the performance and fairness of each classifier, measured by the difference in true positive rates and the classification accuracy respectively. In the bottom right corner we find fair and accurate classifiers. In terms of fairness reweighing and the equalized odds post-processing method perform best. However, the regular random forest classifier comes close to their fairness performance and performs slightly more accurate. The fairness adjusted logistic regression performs worst in terms of accuracy and fairness. As the picture could change depending on the chosen fairness metric (y-axis), we also tried out other metrics, such as equalised odds or predictive parity. In all cases the regular random forest does not perform worse in terms of fairness but better in terms of accuracy than most fairness adjusted classifiers. This supports the in general low racial disparities the fairness audit in the previous section showed.

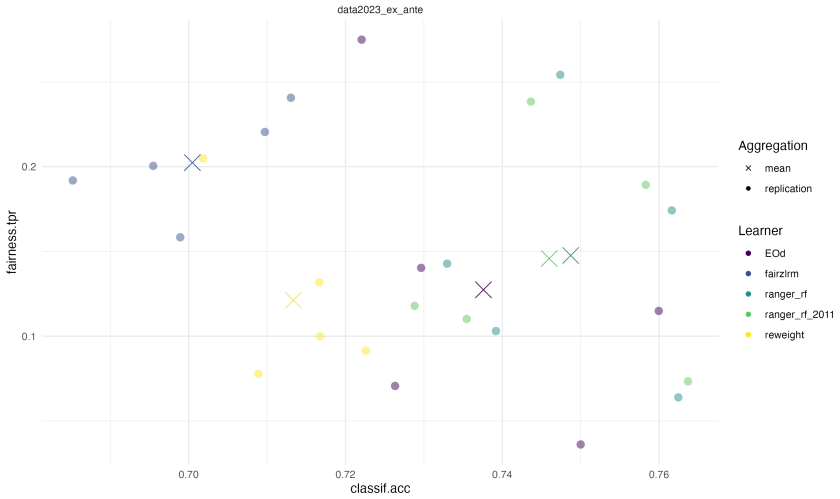

Figure 6: Fairness metrics for the fairness adjusted random forest classifier.

Training on stops from an unconstitutional period

Considering that the SQF practice from 2004 to 2012 has been deemed unconstitutional, the question arises whether a classifier trained on data from this period shows more racial disparities. We choose data from 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 that these are more than 50 times more stops than in 2023. The 2011 data has substantially more low-risk stops, only around 6% of stops result in an arrest. This is a stark contrast to the 2023 data, where 31% of stops resulted in an arrest.
The differences in arrestment rate between groups are slightly lower for 2011 and the highest arrestment rate remains to be for the white group. Due to the large proportion of low-risk stops in 2011 the predicted probabilities are generally low. In fact, the predicted probabilities for the positive class are generally so low that the classifier rarely predicts the positive class for any person, regardless of group membership. A classifier trained on 2011 data primarily suffers from the highly skewed distribution of arrests. A reweighing technique would first need to establish more balance in the target before any fairness analysis becomes relevant.

The SQF practice is fair?

All in all, it seems like a classifier trained on SQF data to predict the arrest of a suspect is not discriminatory against PoC. In contrast, it even performs better for PoC than for white people. This opposes the public belief that the NYPD Stop-and-Frisk practice is biased towards PoC. Badr and Sharma 2022 have similar findings. In their study they choose six representative machine learning algorithms (Logistic Regression, Random Forest, XG Boost, GNB, SVC) to predict the arrest of a suspect. Fairness is measured with different metrics (Balanced Accuracy, Statistical Parity, Equal Opportunity, Disparate Impact, Avg. Odds Difference, Theil Index) and separate analysis are conducted with sex and race as PA. They compare the fairness of the regular learner to the fairness of

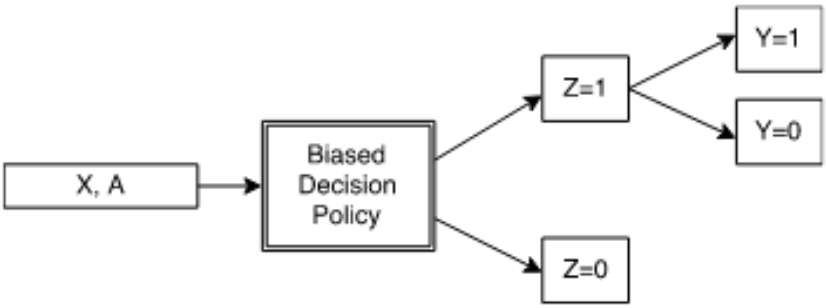

Figure 7: Selection bias in the SQF data.

learner with a pre-processing method (reweighing) and a post-processing method (Reject Option-based Classifier). All in all, they find that the regular models to not perform worse in terms of fairness than the fairness adjusted models. This leads them to conclude "[...] that there is no-to-less racial bias that is present in the NYPD Stop-and-Frisk dataset concerning colored and Hispanic individuals." They further note that the NYPD has made efforts to reduce racial bias in the SQF practice and show the rapid fall in stops in 2011 and the consistently low stop levels since 2014.
Is this the whole picture? Let us look at other studies that approach situation differently.

Sources of bias in the SQF data

To answer the question of fairness in Stop-and-Frisk some other studies take a step back and identify a problem with how the data is generated. In their paper "Residual Unfairness" Kallus and Zhou 2018 conceptualize the problem as shown in Figure 7. 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. This is the first potential source of bias. In the SQF context we can imagine that the police is generally more suspicious towards PoC than white people. Alternatively, we can imagine that they are stopping anyone more likely in high crime areas which happen to be correlated with low-income neighbourhoods which are mostly populated by PoC sources.
Based on this biased decision policy, individuals are either included in the sample or excluded from it $Z \in \{0, 1\}$. Naturally, we can only know the outcome $Y \in \{0, 1\}$ of a stop for the people who were stopped. Kallus and Zhou 2018 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. The indicator $T \in \{0, 1\}$ tell us whether a person belongs to the target population. If $T = 1$ constantly, it means that the algorithm should be deployed for the entire population of NYC.
At this point, we refer back to Figure 1 and Figure 3. Figure 1 shows a clear difference between the racial distribution in the SQF data and the city as a whole. In terms of race, the sample is clearly not representative for NYC. At the same time the estimated

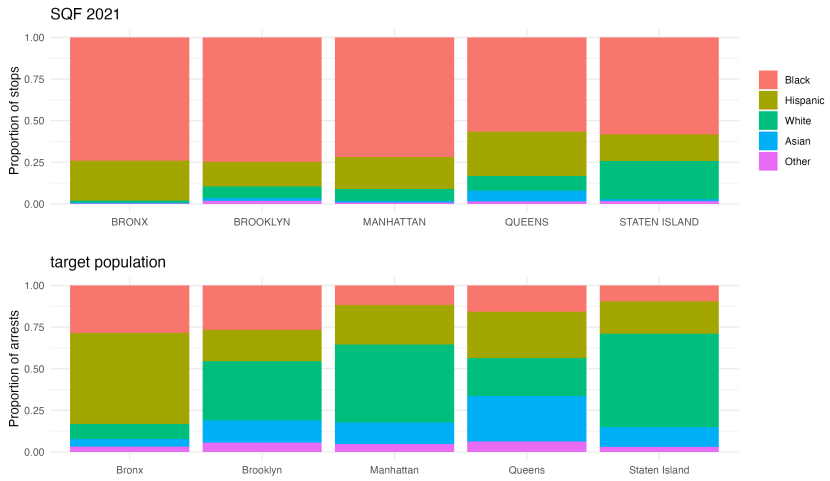

Figure 8: Distribution of race by borough in the SQF data (top) and NYC as a whole (bottom).

borough-specific crime rates also differ from the distribution of stops per borough.

It is unclear whether a biased selection mechanism of such sort, can be captured by group metrics. They work with the joint distribution of $Y, A, \hat{Y}$ and do not take any additional information into account [3] . When we rely on the true label Y to detect unfairness but the true label itself is not reliable (not generated via an objective truth), then the group metrics cannot show this mechanism ()Castelnovo et al. 2022). They offer a rather isolated view on fairness. They assess disparities in algorithmic predictions between protected groups rather than measuring the fairness of a whole situation.
On top of this, the mechanisms behind the selection bias in SQF is twisted in the sense that the (potentially) discriminated group is *more present* in the data. Often we find ourselves in the situation that disadvantaged groups form underrepresented minorities, thus the algorithm oversees them and performs worse on them. In the SQF data, however, the algorithm has plenty of observations from PoC to learn from and less from white people. More training data leads to better algorithmic performance and could explain why the classifier in chapter 5 performed mostly better for PoC. With this in mind Kallus and Zhou 2018 take the following approach.

Alternative approaches to fairness in SQF

In short Kallus and Zhou 2018 come to the conclusion that a SQF-trained classifier exhibits bias towards non-white individuals. Recall that before an officer stops a person, they need to have a suspicion about what the person did wrong. The suspected crime is recorded in the SQF data. The most common suspicion is the illegal possession of a weapon. Kallus and Zhou 2018 limit themselves to only the stops where the suspected

[3] There are variations of group metrics that allow for non-sensitive attributes X to be considered as well when assessing fairness. An example is conditional statistical parity Verma and Rubin 2018.

crime of the illegal possession of a weapon and the goal then becomes to predict the possession of a weapon. We refer to Goel, Rao, and Shroff 2016 for the detailed reasoning behind this approach. Note that this is different to Badr and Sharma 2022, who defined the arrest as the target variable.

Kallus and Zhou 2018 train a logistic regression classifier and measure fairness in terms of equalized odds and equal opportunity. They find that non-white individuals are more often wrongly accused of possessing a weapon than white individuals. They apply a post-processing technique which assigns group specific thresholds to equalize the false negative rates/true positive rates (and the false positive rates/true negative rates in case of equalized odds) that applied group specific thresholds Hardt et al. 2016. After this fairness intervention of course the error rates are equal between groups when tested on the data the algorithm was trained on. However, their contribution lies in designing a way to estimate the error rates that would occur when the algorithm is deployed on the target population. Due to the selection bias in the data, they find that the discrimination against PoC remains in the target population. They call this residual unfairness.
Here I would need to compare the results of one of the fairness classifiers (probably the EoD post-processing) and compare it to the estimated error rates. Does the classifier inherent the same tendencies?

Bias in, bias out - an alternative perspective

An interesting perspective on this observation can be found in Rambachan and Roth 2020. They take a different stance on the problem of biased training data than Kallus and Zhou 2018 and question the "bias in, bias out" mechanism.
They formalise the problem as follows. For the decision-maker (the police) an individual is characterised by the random vector (X, U, A), where X and A have the same meaning as in Kallus and Zhou 2018 and U is a set of unobserved features. These latent variables are unknown to the algorithm but are characteristics the police bases their decision to stop someone on. In the SQF context this could be the personal impression the officer got of a suspect which is not recorded and hard to measure.
The paper assumes the police is a taste-based classifier against African-Americans. This means they hold some form of prejudice against the group of African-Americans that influences their decision to stop a member of this group. In general, for stopping any person, an officer incurs a cost c ¿ 0. If they stop an individual that turns out to be involved in criminal activity and is therefore arrested, the officer receives a reward b = 1 [4]. In case of stopping an innocent person b = 0. For stopping African Americans the payoff an officer expects increases by $\tau > 0$ compared to stopping a white person. The total payoff for stopping an individual is given by:

$$Y + \tau * A - c$$

where Y is the outcome of the stop, τ is the discrimination parameter, $A \in \{0, 1\}$, and $c > 0$ is the cost for stopping a person. Holding the costs c and the outcome of the stop

[4]The reward can set to any number b ¿ 0. We assume b = 1 as in Rambachan and Roth 2020 without loss of generality.

Y constant, searching an African American results in a higher payoff than searching a white person. The goal of the police is to maximise their payoff. Therefore they stop an individual according to the following threshold rule:

$$Z(X, U, R) = 1(E[Y|X, U, A] \geq c - \tau * A)$$

This means that the threshold for stopping an African American is *lower* than for stopping a white person. Consequently, the police stops African Americans more leniently than white people. This taste-based discrimination rule is the biased decision policy introduced in Kallus and Zhou 2018. In Rambachan and Roth 2020 the authors speak of "selective labels" where again the tupel (Y, X, A, Z) is only available for $Z = 1$.
In short: It actually depends on the outcome and the training sample whether the discrminiation of the previously discriminated (bias interhitance) exists. In some cases it can actuall come to the opposite effect, which they call bias reversal. The mechanism is as follows: the historically discriminated groups is very represented in the sample as being included is an act of discrimination itself. This means we have more training data for the disadvantaged group, they resemble the target population more, as they were more leniently included, and thus the classifier generalised better to the disadvanteged group. When we collect more data for the group, we come closer to the target population and our classifier will work better on the target population for the group with more data.
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 many innocent black people in our data. In Kallus and Zhou 2018 this would lead to a lower learned threshold [5] 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 2020 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 2018 bias against a group results in a less representative sample. In Rambachan and Roth 2020 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 2020 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 2018 paper the

[5] 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.

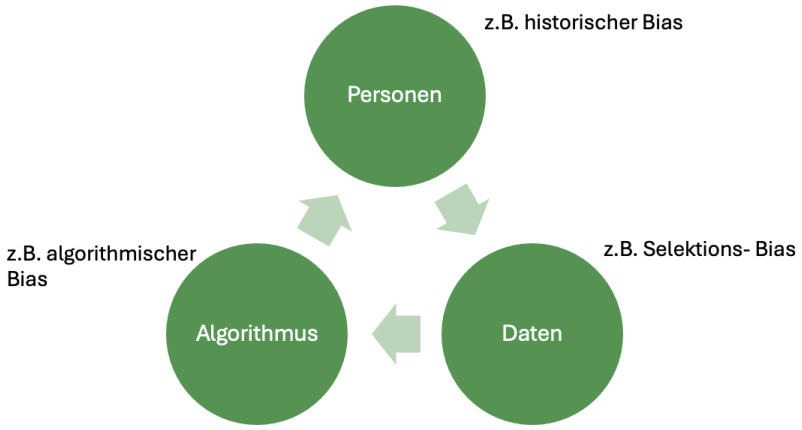

Figure 9: The bias loop.

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), do not 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 2020 we should observe bias reversal.

6 Conclusion

Bias and the feedback loop

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 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 9. At each step of this loop, bias can be introduced in the process and, more dangerous, 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 comes from and how exactly the algorithm will be deployed in practice. With this in mind we can maybe explain they complex and ambigous picture drawn by SQF studies.

Conclusion

In conclusion, the questions of fairness for SQF is difficult. Before any fairness intervention, we have to formulate a clear fairness question. It is something entirely different to ask if the stop, question, and frisk practice (as a whole) is fair or whether a classifier to predict the arrest of a person trained on SQF data is fair? Or whether a classifier trained to predict the possession of a weapon trained on SQF data is fair? The exact question we formulate leads us to look at different aspects of the data. In this paper we got a first idea of the answer to the first questions by comparing certain characteristics of the SQF population to the population of NYC as a whole (descriptive analysis) and find that the two populations do differ. But does it make sense to want the SQF sample be representative for whole NYC or does it not make more sense to want it to be representative of the population of criminals in NYC? Here we see a closer match in racial distributions. This, however, is by far not enough to claim the fairness of the police practice. Crime statistics have to be read with caution. They are influenced by many factors, including the amount of police in a certain area, the socio-economic status of the population and the trust in the police. Historical discrimination leads to lower socio-economic, lower socio-economic status comes with higher crime rates, higher crime rates lead to more police in the area, more police in the area lead to more reported crime. Crime statistics are embedded in a broad context and do not necesarily reflect objective inherent truths but our social and economic system. We can cite Goel, Rao, and Shroff 2016 who approach the question in a more wholeistic way, account for complex factors and come to the conclusion that SQF is over-targetting PoC. As we saw in our own case study and Badr and Sharma 2022 also find, is that this does not mean a classifier trained on SQF data violates group fairness. Depending on the task some classifiers might perform better on the historically disadvantaged group while others in fact discriminate against them. With this study we do not claim to give the answer to fairness in SQF but the goal was to show the readers the complexity of the situation/ give critical perspective/ show different approaches to fairness in SQF. As many datasets, this one comes with a great backstory (socio-economic context, historical biases) and problems (group imbalance, ...) and all of this is entangled. We should be aware of this otherwise it might misinterpretation of results.

List of Figures

List of Tables

Acknowledgement

A Electronic Appendix

Data, code and illustrations are available in electronic form.

References

Badr, Youakim and Rahul Sharma (June 2022). "Data Transparency and Fairness Analysis of the NYPD Stop-and-Frisk Program". In: *Journal of Data and Information Quality* 14.2, pp. 1–14. ISSN: 1936-1955, 1936-1963. DOI: 10.1145/3460533. (Visited on 12/24/2024).

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).

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).

Goel, Sharad, Justin M. Rao, and Ravi Shroff (Mar. 2016). "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy". In: *The Annals of Applied Statistics* 10.1. ISSN: 1932-6157. DOI: 10.1214/15-AOAS897. (Visited on 11/19/2024).

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 (July 2018). "Residual Unfairness in Fair Machine Learning from Prejudiced Data". In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, pp. 2439–2448. (Visited on 12/24/2024).

Khademi, Aria et al. (May 2019). "Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality". In: *The World Wide Web Conference*. San Francisco CA USA: ACM, pp. 2907–2914. ISBN: 978-1-4503-6674-8. DOI: 10.1145/3308558.3313559. (Visited on 12/24/2024).

Kusner, Matt J et al. (2017). "Counterfactual Fairness". In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA. URL: https://papers.nips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Rambachan, Ashesh and Jonathan Roth (2020). *Bias In, Bias Out? Evaluating the Folk Wisdom*.

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, 25th 2025

Juliet Fleischer