Seminar Thesis

FairML and the SQF dataset

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, February, 26th 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

1 Introduction

Building a fair and equitable society is a challenge humans grapple with since ancient times. With the rise of artificial intelligence (AI) questions of justice and fairness have taken on new urgency. AI enables automated decision-making systems (ADM) that are now common in law, healthcare, finance, and other fields, where data-driven decisions can significantly affect lives. Despite their ongoing improvements they carry the risk of perpetuating and even exacerbating social injustices.
After a general introduction to the study of fairness in machine learning (fairML), this paper turns its focus to the stop, question, and frisk (SQF) dataset published by the New York Police Department (NYPD). Since 1990 the US Supreme Court has been allowing police officers in New York City to stop individuals if they suspect them of being involved in criminal activity (Terry v. Ohio (1968), 392 U.S. 1, U.S. Supreme Court.). While proponents argue that SQF is an effective crime prevention tool, many criticize the practice for disproportionally targetting people of colour. The stop-and-frisk practice has a long history of public debate about racial profiling and police trust (see Gelman, Fagan, and Kiss 2007 for more historical details). This makes the datasets recording the stops an interesting resource for fairML research. Advocating for more diversity in the datasets used for fairness research additionally Fabris et al. 2022 recommend the dataset as a valuable resource.
Our main contribution lies in bringing together multiple studies that examine fairness in SQF from different angles. Though these studies seek to answer the same question—Is stop, question, and frisk fair?—they approach the problem differently and arrive at alternative conclusions. This divergence is not necessarily a contradiction but rather a reflection of the diverse perspectives and objectives that shape fairness research. Each study addresses fairness within its own problem setting, making its conclusions valid within that specific context. However, this can create confusion, as studies with different assumptions and goals may still claim to answer the same overarching question. Our goal lies not in identifying *the right* approach, but rather in highlighting the importance of understanding data context and problem framing when evaluating fairness. In Section 3, we introduce the most common fairness metrics and techniques used in machine learning. Next, in Section 4 we apply the theoretical concepts to the real-world SQF dataset. The application on real-world data will show difficulties that come with fairness in practice. This will lead us to explore other studies that have worked with SQF data in Section 5.

2 Related Work

Fairness in machine learning has attracted considerable attention in recent years, leading to a rich literature of definitions and evaluation frameworks. Several works provide broad overviews of these definitions. For example, Verma and Rubin 2018 offers a comprehensive overview of the most popular fairness metrics and Castelnovo et al. 2022 highlights their nuances in a compact manner. Corbett-Davies et al. n.d. and Barocas, Hardt, and Narayanan n.d. serve as detailed resources that offer deeper insights into common falacies in fairness metrics. Beside the definition of fairness a major area of research is the design of bias migitating techniques to which Mehrabi et al. 2022 and Caton and Haas 2024

provide a great overview. Additionally, the `mlr3book` serves as an accessible introduction to the practical implementation of fairness metrics.
Beyond these general discussions, a number of studies have focused on the stop, question, and frisk (SQF) dataset—an area where machine learning intersects with socio-economic and statistical analysis. In this context, Gelman, Fagan, and Kiss 2007 has been one of the earlier works that provides both historical context and a sophisticated statistical to show racial disparities in the policing strategy. Building on this, Goel, Rao, and Shroff 2016 advance the statistical methods further to support the claim that non-white individuals are disproportionately targeted by the New York police.
Moreover, fairness in SQF has also been examined from a causal perspective. For instance, Khademi et al. 2019 explores causal individual and group fairness and introduces causal group fairness metrics for SQF. Their study supports the complexity of measuring fairness in SQF practice as their different metrics come to divergent fairness conclusions. The other studies on SQF Badr and Sharma 2022; Rambachan and Roth n.d.; Kallus and Zhou 2018 will be more closely examined in the final chapter of this paper.

3 Fairness Metrics and Methods

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. What all of them have in common is that they build on the idea of a protected attribute (PA) or alternatively called sensitive attribute. This is a feature present in the training data because of which individuals should not experience discrimination. Examples for sensitive attributes are race, sex and age. Often they are protected by law in some form. Coming back to differences in fairness definitions, 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. Group membership is encoded by the PA. 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 these structures. 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 PA. For this work, we will also stay within this setting. For the subsequent sections let $Y \in \{0, 1\}$ be the true label, $\hat{Y} \in \{0, 1\}$ be the prediction label, $\hat{R}$ bet the prediction score, A be the sensitive attribute and X encode the non-sensitive attributes.

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

Table 1: Group fairness metrics

3.1 Group fairness

The groups metrics presented in the following are observational metrics. They can be separated into three main categories shown in Table 1, depending on which information they use.

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. 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

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

Conditional statistical parity is an extention of this as it allows to condition on A and a set of legitimate features E. In the context of SQF, 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 categories of group fairness metrics can both be derived from the error matrix.

Separation

Separation requires independence between $\hat{Y}$ and A conditioned on the true label 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

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

- Predictive equality/ False positive error rate balance

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

- 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. Thus to formulate equal opportunity one can equivalently require equal trupe positive rates between groups. 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. We note that there is dedicated work that assists in finding the right fairness metric for a given situation and refer to Makhlouf, Zhioua, and Palamidessi 2021 for an alternative approach and more depth.

Sufficiency

Sufficiency requires independence between Y and A conditioned on $\hat{Y}$. 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:

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

Just a for the *Separation* metrics one can combine two of these *Sufficiency* metrics and require them to hold simultaneously to get a stricter requirement. While it is easy to get lost by the amount of fairness definitions in the beginning, 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 evidently follow the same pattern as the other metrics.

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

Table 2: Confusion matrix

Score-based fairness 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. See Verma and Rubin 2018 for an example.

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.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. 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

[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.

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. n.d.) 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 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.
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.

Fairness methods

Another question fair machine learning deals with is how algorithms can be adjusted so that they fulfil one of the above fairness metrics. Depending on when they take place in the machine learning pipeline, we distinguish between

1. Pre-processing methods
2. In-processing methods
3. Post-processing methods

Pre-processing methods follow 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. The idea is to assign different weights to the observations based on relative frequencies, so that the algorithm learns on a balanced dataset Caton and Haas 2024. In-Processing methods modify the optimization criterion, such that it also accounts for a chosen fairness metric. Introducing a regularization term to the loss function is one example of such modifications. Post-processing 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 fulfilled. One example for this is thresholding, where we set group specific thresholds to re-classify the data after training Hardt et al. 2016. Maybe mention some mlr3fainress implemented methods or refer to an overview.

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 1. 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. This framework should be kept in mind when implementing fairness in the real world and it will also help to think about the following case study.

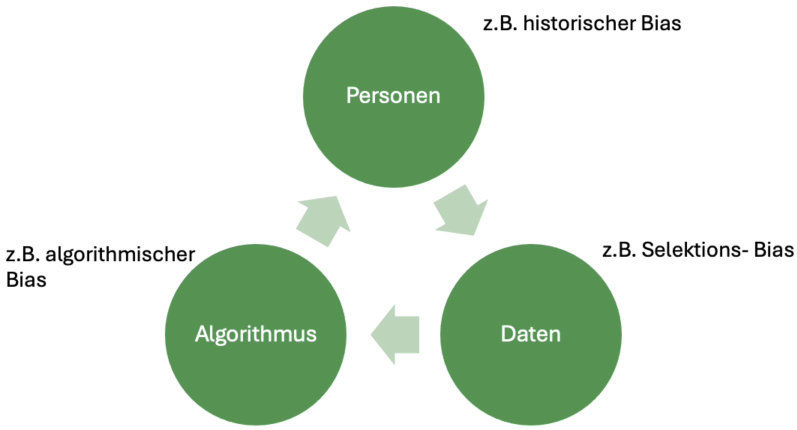

Figure 1: The bias loop.

4 Case Study: Stop, Question, and Frisk

Stop, Question, and Frisk data

After introducing the theoretical tools for assessing fairness, we turn to a case study on the stop-and-frisk practice in NYC. 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. As mentioned in the introduction the so-called "New York strategy" Gelman, Fagan, and Kiss 2007 is highly controversial. 2013 the stop-and-frisk practice during 2004 to 2012 has been deemed as unconstitutional, violating the fourth and fourteenth amendment Source

4.1 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

[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.

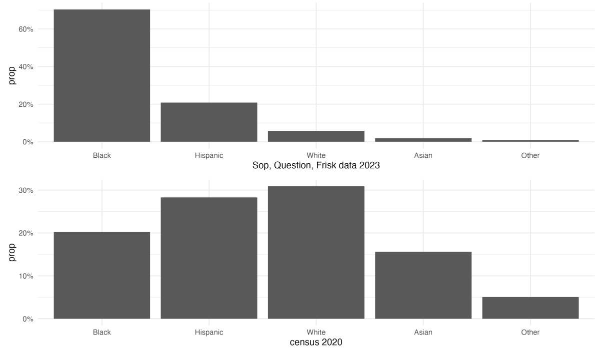

Figure 2: Comparison of race distribution in the training and target population.

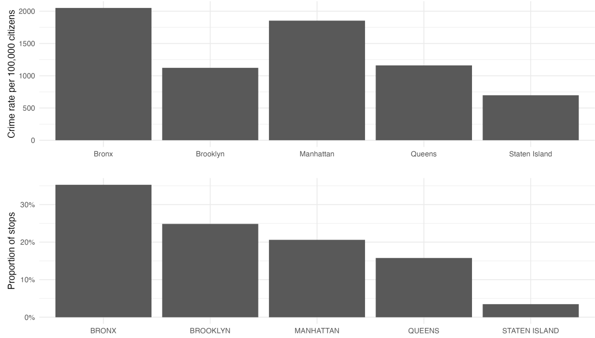

Figure 3: Estimated borough-wise crime rates (top) versus the proportion of stops in each borough (bottom).

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 2. 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.

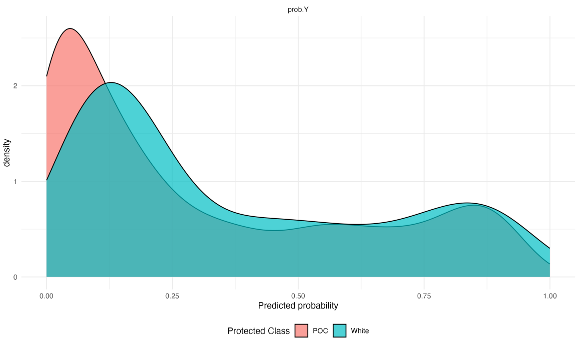

Figure 4: Density of predicted probabilities for both groups.

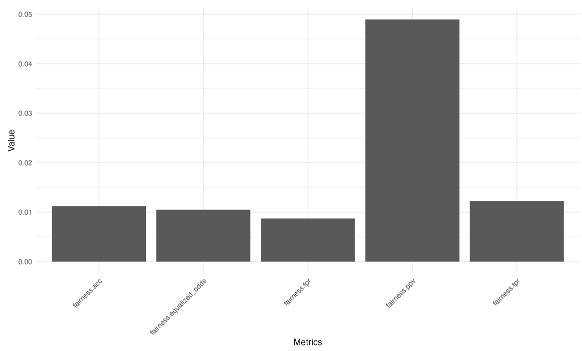

Figure 5: Another relevant plot.

4.2 Fairness Audit

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 they made the decision to arrest the person. This includes information about the development of the stop, e.g. whether the person was frisked or a summon issued (we assume that all of these constitute "smaller" hits that happen before an officer chooses the most extreme consequence, an arrest). A full list of the variables in our model can be found in the appendix. 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 (Badr and Sharma 2022). 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. The mode for the scores for non-white individuals is around 0.05 while it is around 0.125 for white individuals. The score resembles the probability of being predicted positive (arrested). In Figure 5 we plot the absolute difference in selected 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 each of the selected metrics, though the difference in positive predictive rates is close to 0.05. For a more nuanced picture, we additionally report the group-wise error metrics in Table 3. The true positive rate, false positive rate, and the accuracy is basically identical between the two groups. So the Separation metrics are fulfilled. More ore less notable differences can only be seen in the Sufficiency metrics: the negative predictive values/ positive predictive value.

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 on many of the common performance metrics for PoC than for white people. This opposes the public

Metric	PoC	White
tpr	0.75	0.74
npv	0.89	0.85
fpr	0.07	0.06
ppv	0.84	0.89
fdr	0.16	0.11
acc	0.88	0.86

Table 3: Groupwise Fairness Metrics (2023)

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 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." What both of our case studies have in common is that the models were trained on recent data. We trained our model on 2023 stops and Badr and Sharma 2022 used 2019 stops. Since the judgement of how stop-and-frisk was implemented in NYC in 2013, the number of stops has decreased significantly and citizens are in generally less often stopped. After 2014 the stops have been consistently kept at a low level. Badr and Sharma 2022 see this as explanation for their results and state "The NYPD has taken crucial steps over the past years and significantly reduced racial and genderbased bias in the stops leading to arrests. This conclusion nullifies the common belief that the NYPD Stop-and-Frisk program is biased toward colored and Hispanic individuals."

Training on stops from an unconstitutional period

Naturally, the question arises how a classifier trained on data from the unconstitutional period performs. 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.

In the data, 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 low prevalence in the population, a classifier trained on 2011 data primarily suffers from the highly skewed distribution of arrests. In terms of fairness, it performs similar to the 2023 classifier. So somehow the picture is confusing. A classifier trained on SQF data from recent years seems to be fair, and a classifier trained on data from the unconstitutional period is even fairer. In the following we examine alternative approaches to fairness in SQF to find possible a explanation.

5 Studies on the SQF Dataset

5.1 Sources of bias in the SQF data

The main difficulty is that the SQF data contains many information and various tasks could be defined for it. But only some of them are suitable to make conclusions about the fairness of the stop-and-frisk policy as a whole. AS Badr and Sharma 2022 we trained a classifier to predict arrest and used group metrics to assess fairness. Given that our 2011 classifier performed well on the gorup metrics, but the stop-and-frisk practice was officially declared unconstitutional for 2011, fairness measured with these metrics for this classification task is not a good indicator for the fairness of the stop-and-frisk policy.
So to answer the question of fairness in Stop-and-Frisk other studies take a step back and identify a problem with how the data is generated. They try to formalize and acknowledge that the discrimination in SQF does not lie in the outcome of the stop but the decision to stop someone in the first place. In their paper "Residual Unfairness" Kallus and Zhou 2018 conceptualize the problem as shown in Figure 6. 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 2 and Figure 3. Figure 2 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 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

[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.

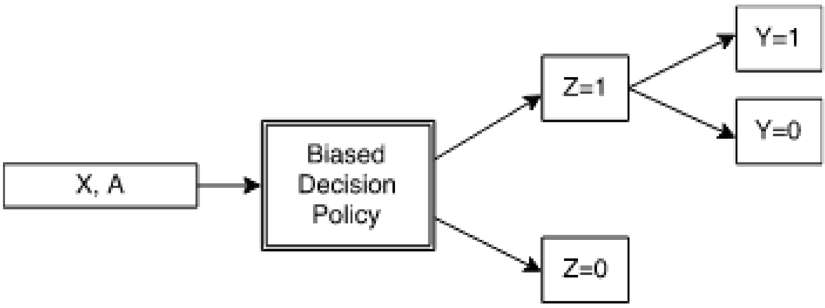

Figure 6: Selection bias in the SQF data.

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.

5.2 Different approaches to fairness in SQF

Residual unfairness

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 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 and our analysis, where the arrest is defined 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) (Hardt et al. 2016). After this fairness intervention the error rates are equal between groups when tested on the data the algorithm was trained on. However, when they claim that when the fairness-adjusted algorithm would be deployed on the

population of NYC as a whole, racial discrimination against the historically discriminated persists. They do not test their dataset on actual new data from citizens but they design a way to estimate the error rates that would occur when the algorithm is deployed on the target population. They call the unfairness that comes from switching from training population to target population **residual unfairness** and identify it when training a classifier to predict the possession of a weapon on SQF data.
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?

Another perspective is offered by Rambachan and Roth n.d. While the main message of Kallus and Zhou 2018 is that even fairness adjusted classifiers exhibit the "bias in, bias out" mechanism Rambachan and Roth n.d. argue that is depends on the chosen classification task.
Similar to Kallus and Zhou 2018 they are interested in whether a person carries a contraband. 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. More precisely, they see the biased-decision policy in the decision to search someone; only on searched people a contraband can be found. Essentially the problem setting is the same as in Kallus and Zhou 2018.

They formalise the problem as follows. For the decision-maker (the police) an individual is characterized 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. In general, for searching any person, an officer incurs a cost $c > 0$. If they search an individual that truly carries a contraband, the officer receives a reward $b = 1$ [4]. In case of searching 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 search, τ is the discrimination parameter, $A \in \{0, 1\}$, and $c > 0$ is the cost for searching a person. Holding the costs c and the outcome of the search Y constant, searching an African American results in a higher payoff than searching a white person. The goal of the police is to maximize their payoff. Therefore, they search an individual according to the following threshold rule:

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

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

This means that the threshold for searching an African American is *lower* than for a white person. Consequently, the police searches African Americans more leniently than white people. In Rambachan and Roth n.d. the authors speak of "selective labels" where again the tupel (Y, X, A, Z) is only available for $Z = 1$.

Given this biased-selection mechanism that produces the training data, the authors distinguish between three classification scenarios.
In the first one, the goal is to predict the possession of a contraband, but the algorithm is trained on the biased sample that searched African Americans more leniently than white people. In this case the algorithm will exhibit *less* bias towards African Americans in the future. What happens is that as the police becomes more biased towards African Americans, they search them more leniently. This means that many innocent African Americans are included in the searched observations. Consequently, the model learns on average lower risk scores for African Americans. Essentially, the data for African Americans becomes more "noisy", more innocent people, without contraband are included, lowering the predicted probabilities for this group. The authors call this mechanism **bias reversal**. In their second classification task the idea is to train an algorithm to predict whether to search someone in the first place. Now the search becomes the target and here bias inheritance is observed. The same goes for a two stage classification ask that first predicts whether to search and then whether the individual carries a contraband if they were predicted as searched. What happens here is that as more of the stopped African Americans are also searched, the algorithm learns to associate search with more with African Americans than with white people and thus in the future also predicts higher probabilities for a search for African Americans. This is the **bias inheritance** mechanism. We can see parallels to this paper to our own case study in the sense that, PoC indeed have lower risk scores (density plot) and are relatively speaking less often predicted as arrested as white individuals.

6 Conclusion

More concrete limitations and what future work could adress. Be as conrete as possible. 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).

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (n.d.). "Fairness and Machine Learning". In: ().

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: ().

Fabris, Alessandro et al. (Sept. 2022). "Algorithmic fairness datasets: the story so far". In: *Data Mining and Knowledge Discovery* 36.6, pp. 2074–2152. DOI: 10.1007/s10618-022-00854-z. URL: https://doi.org/10.1007%2Fs10618-022-00854-z.

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).

Gelman, Andrew, Jeffrey Fagan, and Alex Kiss (Sept. 2007). "An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias". In: *Journal of the American Statistical Association* 102.479, pp. 813–823. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214506000001040. (Visited on 01/08/2025).

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. (n.d.). "Counterfactual Fairness". In: ().

Makhlouf, Karima, Sami Zhioua, and Catuscia Palamidessi (May 2021). "On the Applicability of Machine Learning Fairness Notions". In: *ACM SIGKDD Explorations Newsletter* 23.1, pp. 14–23. ISSN: 1931-0145, 1931-0153. DOI: 10.1145/3468507.3468511. (Visited on 12/01/2024).

Mehrabi, Ninareh et al. (July 2022). "A Survey on Bias and Fairness in Machine Learning". In: *ACM Computing Surveys* 54.6, pp. 1–35. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3457607. (Visited on 01/07/2025).

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, 26th 2025

Juliet Fleischer