Seminar Thesis

FairML and the SQF Dataset

Department of Statistics
Ludwig-Maximilians-Universität München

Juliet Fleischer

Munich, February, 27th 2025

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

Abstract

With the increased presence of AI in our society topics of social justice and fairness have swept over to technical research fileds. 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 how it affects algorithmic learning. For this, we turn our focus to studies that have worked with the SQF dataset and established interesting theoretical results: residual unfairness and bias reversal. The main contribution of this paper lies in comparing and contrasting the different ways in which fairness has been studied for the Stop, Question, and Frisk dataset. We show that the challenges is not to identify the right approach, but rather to understand the implications and reasoning behind each method.

Contents

1 Introduction

The challenge of creating a fair and equitable society has been a concern for society 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 they can affect the lives of people significantly. 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 yearly 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 have a reasonable suspicion that they are involved in criminal activity (Terry v. Ohio (1968), 392 U.S. 1, U.S. Supreme Court).

While proponents argue that the stop-and-frisk strategy is an effective crime prevention tool, many criticize the police for disproportionally targetting people of colour (PoC). The way in which the stop-and-frisk practice was being implemented during 2004 to 2012 in NYC was indeed deemed unconstitutional in 2013, violating the fourth and fourteenth amendment Source. Official statistics show the steep decline in stops after the 2013 judgement. Since then the stops have been kept at a low level.[1] More details on the historical background regarding the public debate about the policing strategy can be found in Gelman, Fagan, and Kiss 2007.

Disagreement on the justice of this system can also be found in the literature. The main contribution of this thesis lies in reviewing multiple studies that examine the fairness of 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.

The paper is organized in the following way: 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.

[1] https://www.nyclu.org/data/stop-and-frisk-data

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 offer a comprehensive overview of the most popular fairness metrics, accompanied by a case study on the Adult dataset. Castelnovo et al. 2022 highlights their nuances and subtleties. The work of Corbett-Davies et al. n.d. and of Barocas, Hardt, and Narayanan n.d. serves as detailed resources that offer deeper insights into common fallacies in fairML.
Beside the definition of fairness a major branch of research has concerned itself with the design of bias migitation techniques. Mehrabi et al. 2022 and Caton and Haas 2024 provide a detailed review. Additionally, the fairness chapter in the Pfisterer 2024 serves as an accessible introduction to the practical implementation of fairness metrics and methods.
Beyond these more general works, a number of studies from the fields fairML, Statistics, and Economics, have focused on the SQF dataset. Building on the previously mentioned work of Gelman, Fagan, and Kiss 2007, Goel, Rao, and Shroff 2016 advance the statistical methods used in the former study to further to support the claim that non-white individuals are disproportionately targeted by New York's police. Khademi et al. 2019 examined fairness in SQF from a causal perspective. Their study supports the complexity of this topic as they arrive at divergent conclusions, depending on which of their metrics they use.
In the course of this paper it will become clearer that selection bias is a major concern for the SQF data. The effects of selection bias on fairness and potential ways to counteract them have been studied by Lakkaraju et al. 2017 and Favier et al. 2023. The other studies that explicitly use 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

It is easy to get overwhelmed by the sheer amount of fairness definitions in machine learning. This chapter groups the metrics in an intuitive way and motivate them in the hope to bring some clarity to the readers. What all the metrics 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. Fairness metrics can be classified 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

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

Table 1: Group fairness metrics

with defining fairness for a binary classification task and one, often dichotomized, PA. For this work, we will also stay within this setting. Moreover, our focus will lie on the observational metrics, as causal notions of fairness require more invovled techniques that are out of the scope of this paper.
We denote the categorical sensitive attribute as $A \in \{a, b\}$ while we assume for simplicity that it is binary. The remaining features are encoded as $X \in \mathcal{X}$.
We define $f : \mathcal{X} \times \{0, 1\} \to [0, 1]$ as a prediction function that returns a score $s = f(x, a)$ representing the estimated probability that the true label is 1 for each input (x, a) .
To obtain a hard prediction from the score, we define a thresholding function

$$g(s) = \begin{cases} 1, & \text{if } s \geq c, \\ 0, & \text{if } s < c, \end{cases}$$

where $c \in [0, 1]$ is a predetermined threshold (often $c = 0.5$). Thus, the final predicted label is given by $\hat{y} = g\big(f(x, a)\big) = \mathbf{1}\{f(x, a) \geq c\}$
To facilitate the understanding of the following fairness metrics, we can choose variables from the SQF dataset for illustration.

3.1 Group fairness

They observational group metrics presented in this section can be separated into the 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. For a binary classification task with binary sensitive attribute this can be formalized as

- *Demographic parity* requires equal positive prediction ratios (ppr) for both groups

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

Independence is best, when a form of equality between groups should be enforced, 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?.
In many cases it can make sense to allow for additional information to be taken into account. Therefore an extension of demographic parity can be defined as

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

Table 2: Confusion matrix

- *Conditional statistical parity* $P(\hat{Y} = 1 \mid E = e, A = a) = P(\hat{Y} = 1 \mid E = e, A = b)$

E is a set of legitimate features that encapsulates valuable information about predicting the target Y. In the context of SQF, predictive parity could mean that we require that PoC relatively measured are as often predicted positive as white people, regardless of any information. Conditional statistical parity, on the other hand, does not require equal proportions in general but only within specific subgroups (defined via E). For example, we could require equal ppr between PoC and white people who *live within the same borough* of New York ($E = borough$).
The other two categories of group fairness metrics can both be derived from the confusion matrix Table 2.

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* requires the false negative rates, the ratio of actual positive people that were wrongly predicted as negative, is equal between groups

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

- *Predictive equality/ False positive error rate balance* follows same principle as equal opportunity but for the false positives

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

- *Equalized odds* requires that both the true positive rate and the false positive rate are equal across groups

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

- *Overall accuracy equality* requires equal accuracy for both groups

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

- *Treatment equality* builds groups-wise ratios of error-rates and requires equality

$$\frac{\text{FN}}{\text{FP}}\Big|_{A=a} = \frac{\text{FN}}{\text{FP}}\Big|_{A=b}$$

The idea behind *Separation* metrics is ... As *Separation* criteria condition on the true label Y it is suitable when we can be sure that Y is free from any bias, meaning it was generated via an objectively true process.

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 rates*

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

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. The intuition behind *Sufficiency* is that ... 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.
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.
In essence the group metrics outlined so far do nothing other than picking a performance metrics from the confusion matrix and requiring it to be equal across two (or more) groups. This means that they come with trade-offs just as the usual performance metrics for classifiers do Kleinberg, Mullainathan, and Raghavan 2017. Researchers have shown that if base rates, i.e. the proportions of the positive outcomes of the groups in the population, differ between groups, it is mathematically impossible to equalize all desirable metrics simultaneously Chouldechova 2016. This is also refered to as the Impossibility Theorem.

Score-based fairness metrics

Most (binary) classifiers work with predictions scores $S \in [0, 1]]$ 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.

Choosing the right group metric

Due to the abudance of group metrics alone there have been further studies to assist practitioners choose the right metric. One possibility is to distinguish between punitive and assistive tasks Ghani et al. 2025. For punitive tasks metrics that focus on false positives, such as predictive equality are more relevant. For assistive tasks, such as deciding who receives a welfare, a focus on minimizing the false negative rate could be more relevant. This points to equal opportunity as suitable metric. 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. There is dedicated work that assists in finding the right group fairness metric for a given situation and refer to Makhlouf, Zhioua, and Palamidessi 2021 for an in-depth analysis.

3.2 Individual fairness

Individual metrics shift the focus from comparison *between* groups to comparison *within* groups. 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.
What then, do we measure this for each individual? Get aggregation statistics from this? Explain this further in one or two sentences.

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.[2] Other ways to assess FTU can be found in Verma and Rubin 2018.
A problem blinding has are proxies. Proxies are variables that are strongly correlated with the sensitive attribute. In the presence of such features, it is not enough to mask the information of the sensitive attribute during training because discrimination can persist via these proxies.
For SQF this could 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, important information could be lost by removing too many features (Castelnovo et al. 2022).

Comparison and Summary

Experts debate the incompatibility of group and individuals fairness. It is out of the scope of this paper to discuss this topic, and we simply point out that the sharp line we drew between group and individuals metrics gets softer as a group metric like demographic parity does not only take information from $Y, \hat{Y}, A$ into account but allows for information contained in the non-sensitive features to seep into the fairness assessment (Castelnovo et al. 2022).

[2]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.

Group metrics are certainly easier to understand and apply as most of them are implemented in fairness software packages. For this reason our case study in Section 4 uses group metrics for the fairness assessment.

3.3 Fairness methods

After defining fairness in a mathematical sense, the question arises how a classifier can be modified to satisfy the chosen definition of fairness. This is what fairness methods deal with. Depending on their position 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). Depending on the task (regression, classification) and the model there are highly specified and advanced methods. For the case study in chapter 3, we limit ourselves to methods implemented in the `mlr3fairness` package.

3.4 Bias and the feedback loop

Before applying the theory to real-world data, it remains to introduce different types of biases and the context in which a machine learning model is usually embedded. Deployed as an ADM, the model assists in decisions such as whether someone gets admitted to college, receives a loan or is released from prison. It thereby indirectly contributes to shaping our reality.
Mehrabi et al. 2022 conceptualise the situation in form of the *data, algorithm, and user interaction feedback loop* (Figure 1), which can be understood as follows. We as a society make decisions, which are reflected in our reality. The reality is made measurable by collecting data. The algorithm learns from this data to make an optimal prediction, on which the decision-maker bases their judgement. The new choice will shape our reality again, which reflects in updated data.
At each stage, bias can be introduced into the process. More dangerous, bias can even

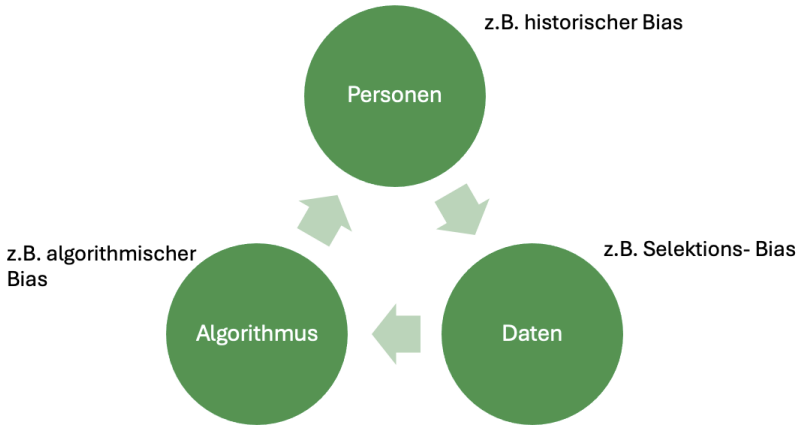

Figure 1: The *data, algorithm, and user interaction feedback loop* as described by Mehrabi et al. 2022. Different categories of bias can be introduced at each stage of the process.

be amplified as the algorithm influences decision-making on a large scale. Consequently, every fairness project comes with the responsibility to understand the data-generating process and gain clarity on how the algorithm will be deployed in the real world.
Moreover, the *data, algorithm, and user interaction feedback loop* helps clarify which type of bias might be relevant in a given situation by placing it at a specific position in the feedback loop. Distinguishing between bias mechanisms can be crucial and should influence the definition of fairness and the choice of fairness adjustments in a given situation. This will also become evident in the following section where we examine the SQF dataset.

4 Case Study: Stop, Question, and Frisk

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) has been criticized for disproportionally targetting African American and Hispanic individuals. This makes the recordings of the stops an interesting resource for fairness research. It also has been recommended by **Fabris˙2022** for fairML studies. Not lastly in the aim to bring more diversity to the datasets used in this field. For our analysis we are interested in whether a classifier trained to predict the arrest after a stop is discriminatory with respect to race.

4.1 Fairness Experiment: Setup

We compare the following models in terms of fairness and model performance, measured by the difference in true positive rates (equal opportunity) and the classification accuracy respectively:

- Regular Random Forest
- Reweighing to balance disparate impact metric (Pre-Processing)
- Classification Fair Logistic Regression With Covariance Constraints Learner (In-Processing)
- Equalized Odds Debiasing (Post-Processing)

More details about the methods can be found in Pfisterer 2024. Specifically, for Reweighing, see mlr3fairness Reweighing. Refer to Fair Logistic Regression for more details on the chosen in-processing method. For the post-processing strategy, check Equalized Odds.

4.2 Data description

As they were the most recent at the time of writing this paper, we work with the stops from 2023. 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.[3]

[3]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 will introduce complications when we implement fairness methods.

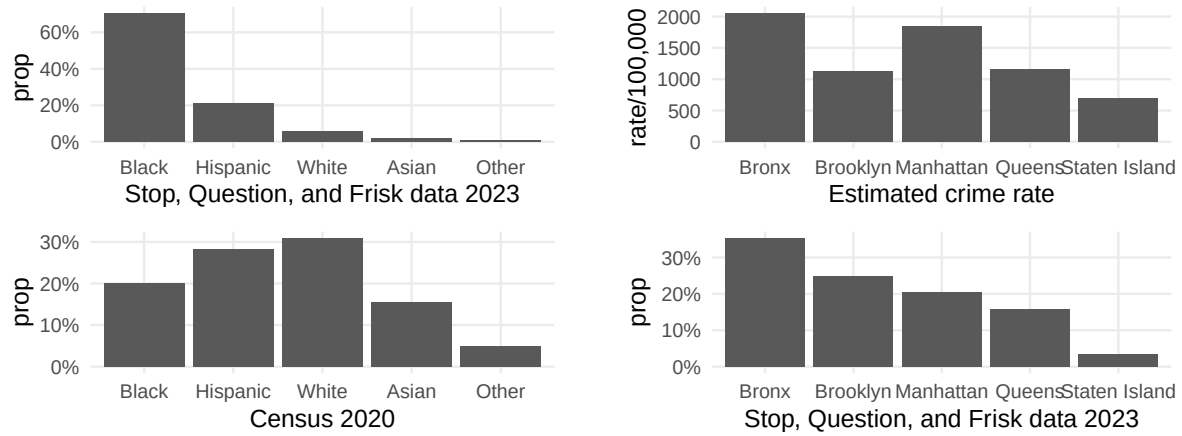

Figure 2: Bar plot comparing the distribution of ethnic groups across boroughs in the SQF 2023 and NYC from 2020 Census (left). On the right a comparison of the estimated borough-wise crime rate per 100,000 citizens with the ethnic distribution of SQF stops.

We 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 contribute to only 20% of the city's population (Figure 2, left). At the same time white people form the majority of New York citizens (30%) but are involved in merely 6% of the stops. After 2012 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 occured 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 2, right).

After a more general overview of the dataset, we turn to the outcome of interest. In the cleaned 2023 data about 31% of stops result in an arrest. Table 3 shows the the disparities in arrestment across ethinic groups is in general low for 2023. 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.
Given the development of stops over the years and the judgement in 2013[4], the question arises if a classifier trained on data from the unconstitutional period (2004-2012) will perform differently. For a comparison in fairness and performance, we therefore train an additional random forest classifier on data from 2011. This is the year with the most stops. We carry out the same data cleaning steps, starting with 685,724 recorded stops and reducing this to 651,567 clean observations. Note that these are around 40 times more stops than in 2023. This means that the 2011 data has substantially more low-risk

[4]https://www.nyclu.org/data/stop-and-frisk-data

Group	Prop
Black	30.56%
Hispanic	31.60%
White	37.95%
Asian	37.84%
Other	31.45%

Table 3: Groupwise Arrestment Rates in 2023

Group	Prop
Black	5.988%
Hispanic	5.830%
White	6.859%
Asian	5.840%
Other	4.575%

Table 4: Groupwise Arrestment Rates in 2011

stops; only around 6% result in an arrest. This is a stark contrast to the 31% in 2023. 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 people.

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. 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 Badr and Sharma 2022.

4.3 Results of the Fairness Experiment

For the training of the classifiers, we dichotomize the race attribute by grouping "Black" and "Hispanic" as people of colour ("PoC") and "White", "Asian", and "Other" as white ("White"). We run a five-fold cross validation and show the results in Figure 3. 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. Somewhat surprisingly, it does not make any difference for the fairness if the classifier is trained on 2011 or 2023 data. We examined the model closer and find that due to the low prevalence in the population, a classifier trained on 2011 data primarily suffers from the highly skewed distribution of arrests. The classifier largely predicts the negative label for *anyone* regardless of race, which overshadows potential fairness concerns. 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.
Since the classifiers perform similarly, we choose the regular random forest trained on 2023 to examine the model closer. On the left 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 prob-

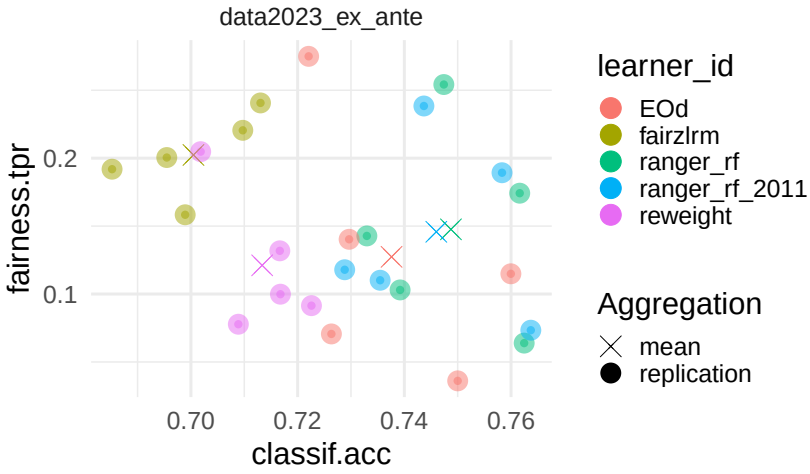

Figure 3: Comparison of learners with respect to classification accuracy (x-axis) and equal opportunity (y-axis) across (dots) and aggregated over (crosses) five folds.

ability of being predicted positive (arrested). On the right Figure 4 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 5. 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.

	TPR	NPV	FPR	PPV	FDR	Acc
PoC	0.75	0.89	0.07	0.84	0.16	0.88
White	0.74	0.85	0.06	0.89	0.11	0.86

Table 5: Groupwise Fairness Metrics (2023)

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. Badr and Sharma 2022 have similar findings.
In their study they choose six representative machine learning algorithms (Logistic Regression, Random Forest, Extreme Gradient Boost, Gaussian Naïve Bayes, Support Vector Classifier) to predict the arrest of a suspect. Fairness is measured with six 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

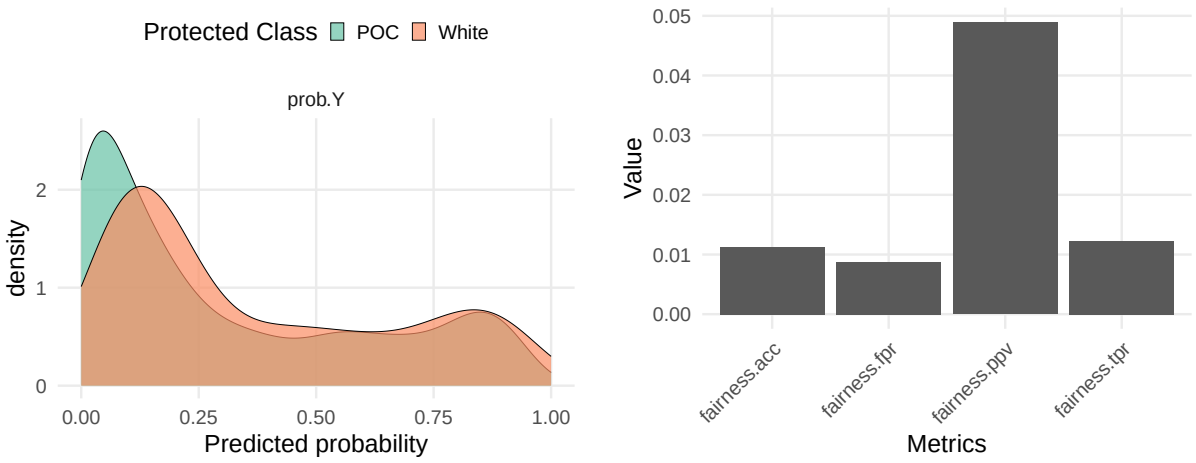

Figure 4: Fairness prediction density plot (left) showing the density of predictions for the positive class split by "PoC" and "White" individuals. The metrics comparison barplot (right) displays the model's absolute differences across the specified metrics.

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." Is this the whole picture?

5 Studies on the SQF Dataset

5.1 Different approaches to fairness in SQF

Before going into detail about a specific study, we provide a tabular overview of the different approaches to fairness in the SQF data. We will go into more depth into some of them in the following.

Authors	Task	Model	Fairness Metric	Results
Kallus and Zhou 2018	Predict prob. of innocence (no weapon)	Log. Regression	Equal Opportunity, Equalized Odds	Bias against PoC
Rambachan and Roth n.d.	Possession of contraband	Log. Regression	No explicit fairness metric; evaluate prediction function properties	No bias against PoC
Badr and Sharma 2022	Predict probability of arrest	Log. Regression, RF, XGBoost, GNB, SVC	Balanced Accuracy, Stat. Parity, Equal Opportunity, Disparate Impact, Theil Index	No bias against PoC
Khademi et al. 2019	Predict probability of arrest	Weighted regression models	FACE causal fairness (group), FACT fairness (individual)	No group bias, but individual bias
Goel, Rao, and Shroff 2016	Predict possession of weapon	(Penalized) Log. Regression	No explicit fairness metric; group-wise hit rates	Bias against Black and Hispanic

Table 6: Summary of SQF-related Fairness Studies

One of the main difficulties that come with the NYPD's data is that, when asking whether stop-and-frisk as a policing strategy is fair, one can come up with various tasks to try to answer this question. 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 the arrest and used group metrics to assess fairness. Given that both, the 2011 and 2023 regular random forest classifier, performed well on the group 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 policy as a whole.
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 formalize and acknowledge that the discrimination in SQF does not solely lie in the outcome of the stop but the decision to stop someone in the first place.

5.2 Residual unfairness

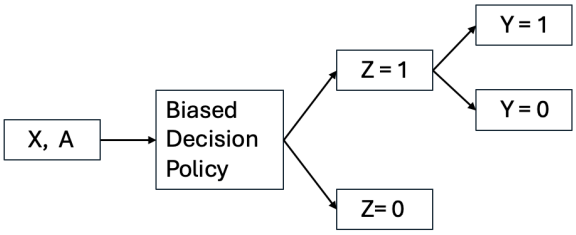

Figure 5: Selection bias in the SQF data, as conceptualized by Kallus and Zhou 2018. The true label is only known for the stopped individuals ($Z = 1$).

In their paper **Residual Unfairness in Fair Machine Learning from Prejudice Data** Kallus and Zhou 2018 conceptualize the problem as shown in Figure 5. A person is defined 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 ($Z = 1$) or not ($Z = 0$). This is the first potential source of bias. It can be seen as a category of selection bias and, referring back to the feedback loop (Figure 1), is introduced by the user.
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 that happen to be mostly low-income neighbourhoods which are largely populated by PoC.
Naturally, we can only know the outcome $Y \in \{0, 1\}$ of a stop for the individuals who were stopped. This can create a situation in which the training data produced by the biased decision policy is not be representative for the population the algorithm will be deployed on.
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 problem for fairness in this case is that fairness adjustments of the learner (trained on the biased sample) do not translate to the target population. Even fairness-adjusted classifiers can discriminate against the same group that has historically faced discrimination Kallus and Zhou 2018. They call this remaining disparities in fairness metrics **residual unfairness**.
At this point, we refer back to Figure 2. It 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[5]. At the same time the estimated borough-specific crime rates also differ from the distribution of stops per borough as seen in **??**.
They show that their theoretical findings can be observed in the SQF data. Their task is to predict the innocence of a person, while they define innocence based on whether

[5]It can be questioned whether it makes sense to require the SQF sample to be representative for the population of NYC. It might make more sense to require that the it is representative of the population of *criminals* in NYC.

someone carried an illegal weapon (guilty) or not (innocent). The reasoning behind this approach is that the discriminated group is the one that was more often wrongly accused of carrying an illegal weapon. Kallus and Zhou 2018 find that non-white people are indeed wrongfully convicted more often. Even after a post-processing strategy to reach equalized odds, the unfairness against PoC persists as the classifier is used on the target population of NYC as a whole.

5.3 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 $Y \in \{0, 1\}$. 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 $Z = 1$ or not $Z = 0$. Again, only on searched people a contraband can be found. So we are essentially in the same problem setting as before. The goal is to estimate the possession of a contraband Y, but we estimate this from $Y|Z = 1$
In contrast to Kallus and Zhou 2018, the authors of this paper argue that the classifier shows the opposite effect; instead of continuing to discriminate the previously disadvantaged group, the classifier exhibits *less* bias as the prejudice against African Americans increases.
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, which lowers the predicted probabilities for this group. The authors call this mechanism **bias reversal**.

As seen in Table 6 these two studies there are more studies that have worked with the SQF data, each with a unique approach to the question of fairness of the policing strategy. Khademi et al. 2019 are also interested in whether the decision to arrest an individual after a stop has been made is discriminatory with respect to race. They design two causal fairness methods, namely the Fair on Average Causal Effect (FACE) and the Fair on Average Causal Effect on the Treated (FACT), to estimate the causal impact of race on the outcome. While one of their metrics finds that the odds of being arrested after a stop are higher for Black-Hispanics than for white individuals, the other metric does not show any racial discrimination.
Goel, Rao, and Shroff 2016 on the other hand focus on the prediction of the possession of a weapon. They find that Black and Hispanic individuals are disproportionately involved in low-risk stops.

6 Conclusion

Certainly the SQF data comes with interesting questions and challenges. We specifically examined fairness and selection bias, but there are more aspects to explore. Historical bias could also play a role and it would be interesting to see how future studies could incorporate this. Moreover, the impact of the class imbalance in the protected attribute on the fairness of the model could be further investigated. The dataset was rightfully recommended by x, offering research possibilities for various disciplines.
After a detailed fairness audit, a fairness experiment with various learner and the review of multiple studies our answer to "Is the stop, question, and frisk practice fair?" remains to be: it is complex.
With our work we showed that, before any fairness intervention, it is crucial to formulate a concrete 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 the historical stops is fair.
The question we formulate can lead to the design of completely different algorithmic tasks and fairness analysis.

List of Figures

List of Tables

A Electronic Appendix

See the GitHub repository for data, code and illustrations: SQF Fairness Project

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).

Chouldechova, Alexandra (2016). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5 2, pp. 153–163. URL: https://api.semanticscholar.org/CorpusID:1443041.

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

Favier, Marco et al. (Dec. 2023). "How to Be Fair? A Study of Label and Selection Bias". In: *Machine Learning* 112.12, pp. 5081–5104. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-023-06401-1. (Visited on 02/05/2025).

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).

Ghani, Rayid et al. (2025). *Chapter 11: Bias and Fairness — Big Data and Social Science*. (Visited on 01/15/2025).

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).

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2017). "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *LIPIcs, Volume 67, ITCS 2017* 67.

A.1 SeriesInformation

LIPIcs, Vol. 67, 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), pages 43:1-43:23, 43:1–43:23. ISSN: 1868-8969. DOI: 10.4230/LIPICS.ITCS.2017.43. (Visited on 02/27/2025).

Lakkaraju, Himabindu et al. (Aug. 2017). "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM, pp. 275–284. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098066. (Visited on 12/25/2024).

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).

Pfisterer, Florian (2024). "Algorithmic Fairness". In: *Applied Machine Learning Using mlr3 in R*. Ed. by Bernd Bischl et al. CRC Press. URL: https://mlr3book.mlr-org.com/algorithmic_fairness.html.

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

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

Declaration of authorship

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

Munich, February, 27th 2025

Juliet Fleischer