BIAS ANALYSIS ALGORITHM ALPHA

[redacted]

Werkplaats Datagestuurde Aanpak Ondermijning | Gemeente Amsterdam

Introduction **2**

Importance of bias analysis and our goal **3**

Pre-required knowledge **4**
- Bias in machine learning models 4
- Disparate treatment vs. disparate impact 4
- Individual vs. group fairness 5
- Direct vs. indirect bias 5
- How to reduce bias 6

Pre analysis decisions **7**
- What to analyze? 7
- Which metrics to use? 8
- How to do the analysis? 9

Alpha algorithm Bias Analysis **10**
- Preliminary analysis 10
- Direct bias analysis 11
- Indirect bias analysis 13

Results and conclusion **18**

Introduction

Amsterdam has limited living space, both for citizens and visitors. Therefore, when a citizen wants to rent out their home or houseboat to tourists, specific requirements need to be met. For example, there is a maximum of 30 nights per year and a maximum of 4 people at a time, and they must report it to the municipality.

Reports can be made when there is a suspicion that one of those requirements is not fulfilled, for instance by neighbors or rental platforms. If such a report is received, employees of the department of Surveillance & Enforcement can start an investigation.

The *Alpha algorithm* supports the employees of the department of Surveillance & Enforcement in their investigation by prioritizing the reports so that the limited enforcement capacity can be used efficiently and effectively. This goal is achieved by analyzing large amounts of data and calculating the statistical probability of illegal holiday rental at an address. This probability is added to the relevant case in the form of a percentage.

The risk indication of an illegal holiday rental is only one of the factors considered to determine whether an address has a chance of a field service investigation. The stage, the details of the case, requests from enforcers or project employees will, of course, be taken into consideration.

The prediction, and the reason behind why this prediction has been made, are shown to the employee, providing an insight into what the Alpha algorithm has based the risk assessment on.

Whether there is an illegal holiday rental or not, is determined by the responsible supervisors in collaboration with the project enforcers. An administrative and field service investigation is performed, and the case is then discussed intensively in a debriefing session, to be able to make a decision. The Alpha algorithm, therefore, does not make independent decisions.

The purpose of the Alpha algorithm is to support the business in "prioritizing and selecting" cases, and humans will always stay in the lead.

This document describes the process and the outcome of the bias analysis over the Alpha algorithm, performed by the Datagestuurde Aanpak Ondermijning team and requested by the department of Surveillance & Enforcement.

Importance of bias analysis and our goal

Decades of research have compared the accuracy of statistical models to the judgments of humans and found that in many situations, data-driven decisions equalize or improve those based on intuition or expertise.

Machine learning promises to achieve a step further, because it offers to uncover relevant factors that humans might overlook, given the complexity or subtlety of the relationships in historical evidence. To achieve this, a machine learning algorithm learns through the generalization of examples, defining rules that effectively apply for past cases, but also for future unseen cases.

This way of learning by no means ensures that it will lead to fair decisions, especially when using machine learning to model human behavior and its characteristics. Our historical examples of the relevant outcomes will almost always reflect historical prejudices against certain social groups, prevailing cultural stereotypes, and existing demographic inequalities. Finding patterns in this data will often mean replicating these very same dynamics.

Our goal is to identify the bias generated by these patterns in order to train a model that mitigates harmful, unjustified, and unacceptable disparities.

To achieve this goal, it is necessary to estimate the potential magnitude and direction of biases introduced by the model, understanding them, and mitigating when necessary.

Pre-required knowledge

Bias in machine learning models

Three were the main challenges addressed in our analysis:

1. Discover unobserved differences in performance.
2. Cope with observed differences in performance.
3. Understand the causes of disparities in predicted outcomes.

To understand the reasons that may introduce bias in the data, it is useful to look at some types of bias (by no means an exhaustive list):

- Selection bias: when some subjects are more likely to be sampled compared to others.
  An example is when the sample consists of investigations, and people with certain characteristics were more likely to be investigated. The problem is that a model containing this bias may influence the selection of its future dataset, increasing the discrimination.
- Self-selection bias: when subjects can influence whether or not they are included in the sample.
  An example is when a person can decide not to participate in a study.
- Attrition/survivorship bias: when the final sample consists of subjects that "survived" a certain trajectory and excludes those that dropped out.
  An example could be the students that passed a specific course.
- Observer/societal bias: when the data is based on someone's subjective judgment of a situation, it may be biased after the observer's personal biases.

Disparate treatment vs. disparate impact

Two principal doctrines exist in discrimination law: disparate treatment and disparate impact.

Disparate treatment happens when the treatment of an individual is less favorable than treatment of others for discriminatory reasons (as race, religion, national origin, sex, or disability).

It is introduced into the model by including one of the discriminatory characteristics as a feature. It can be formal (including the feature explicitly) or intentional (using the feature, but indirectly or hidden).

It is important to prevent this to achieve procedural fairness and equality of opportunity: so that people with the same characteristics should get the same outcome.

Disparate impact refers to the result of the application of a standard, requirement, test, or other screening tool used for selection that, though appearing neutral, hurts individuals who belong to a legally protected class.

It could be introduced into the model by using features that seem non-discriminating, but they generate discriminatory outcomes nevertheless. Preventing this aims to have distributive justice and minimize inequity: outcomes should be equally divided among groups, even if groups may differ in characteristics.

The logic is that differences in characteristics are caused by historical injustices, so outcomes should be distributed equally to prevent historical injustices from having a lasting impact. It is necessary to be careful, because it could have an adverse impact if, for example, loans are given to people with lower creditworthiness, and they go bankrupt because of it.

To check disparate impact, we will use the "four-fifths" rule: in historical studies, a 20%-difference between two groups has been enough to conclude that there's discrimination.

In summary, the algorithm should either:

- treat similar people similarly based on relevant features, given their current degree of similarity;

or

- treat seemingly dissimilar people similarly, based on the belief that the dissimilarity is a result of past discrimination.

Individual vs. group fairness

Individual and group fairness are closely related to disparate treatment and disparate impact.

Individual fairness states that similar individuals should be treated similarly, whereas **group fairness** states that on a group level, outcomes should be equal.

Note how they are not necessarily compatible: to make the outcomes fair on a group level, we might need to treat individuals from a disadvantaged group less harshly than those from an advantaged group.

Roughly speaking, group fairness is the way to go if one believes that differences in characteristics and outcomes between groups are the result of historical injustices (like discrimination) that an individual cannot do anything about.

Individual fairness is the way to go if one believes that the characteristics and outcomes of a person have nothing to do with their group, but are the sole product of their own choices and actions.

Direct vs. indirect bias

Direct bias means that a model discriminates by using a sensitive attribute, such as group membership, as a feature. The model can thus directly see which group subjects belong to, and directly learn the correlation between the sensitive feature and the target.

The solution suggested to mitigate direct bias is to remove sensitive attributes from the model.

Indirect bias (also called **bias by proxy**) happens when one of the non-sensitive features is correlated with a sensitive attribute. By using this **proxy variable** as a feature, the model may still learn to discriminate between groups.

For example, the postcode is often a proxy for nationality: postcode in itself is not sensitive, but in most cities, postcode contains information about which nationalities you are likely to find in a specific zone.

Indirect bias is also sometimes called "redlining", after the past practice of US credit providers to circle certain neighborhoods on a map with a red pencil to mark them as areas they would not serve. This way, they deliberately reduced the number of black people in their customer base.

How to reduce bias

There are three stages in the modeling process where is possible to intervene and reduce the bias:

- during the preprocessing of the data,
- during the selection/training of the algorithm,
- or as postprocessing of the predictions.

In this project, preprocessing has been chosen to mitigate the bias on the model.

In terms of practical implementation, it is good to remark that many of the metrics and methods mentioned in this document are implemented in the AI Fairness 360 toolkit by IBM.

Preprocessing

The idea behind the preprocessing techniques is to achieve better independence by adjusting the features.

Given a feature X, this is done by finding a new representation X' of X that captures as much information as possible from X, while minimizing the correlation between X' and A (where A is the sensitive attribute).

The technique chosen for this project is **reweighing**, a technique by Kamiran and Calders (2012), which is included in the AIF360 toolkit. An explanation of the technique is going to be omitted since the reweighing has not been necessary for our analysis.

Pre analysis decisions

What to analyze?

An analysis has been run over all the features used by the model, measuring the bias for each of them.

Before starting the analysis, it is important to define which discriminatory categories could be present and should be identified.

The following relevant bias categories have been decided for this project:

- nationality (ethnicity/color/citizenship/family background);
- country of birth;
- gender;
- sex;
- sexual orientation;
- religion;
- age;
- pregnancy;
- familial status;
- disability;
- class (social class, occupation, veteran, etc.);
- skin color;
- ethnicity;
- political view;
- health;
- genetics.

Once the categories are defined, it is necessary to iterate over each feature, brainstorming if they could be directly or indirectly correlated to any of the bias categories.

After assigning the categories to the features, the feature values need to be split into two groups: *privileged* and *unprivileged*. The easiest way to understand how to perform the split is through an example using sex. The most common discrimination for the category sex is against women. So, if the feature **sex** would be present in our dataset, we should split it into:

- privileged: man,
- unprivileged: woman.

In the case of continuous variables, the split is performed by analyzing the distribution of the values and setting a threshold. All the subjects with a value below the threshold will fall in one group, and all the subjects with a value above in the other.

It is important to notice that the bias analysis identifies bias in both groups, so, even if the bias in the previous example would be against men, it will be identified. For this reason, it is possible to discover, after the analysis, that the group we thought was being privileged is unprivileged, and vice versa.

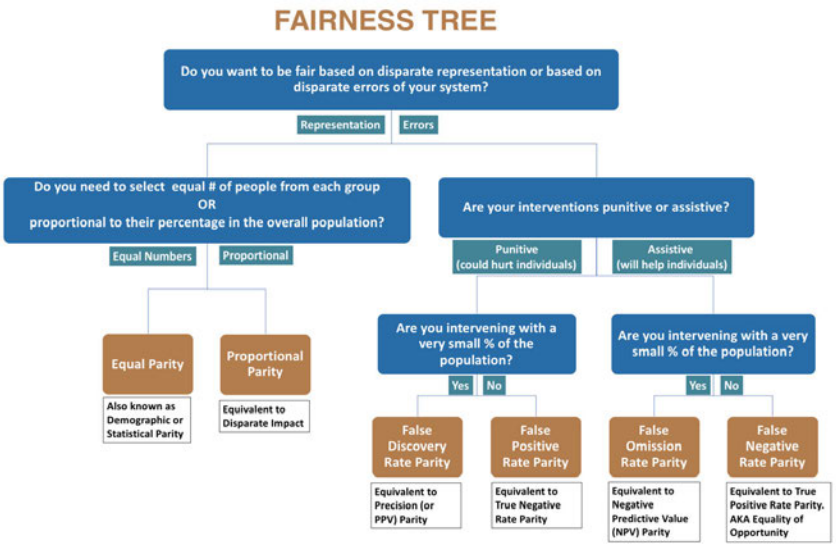
For consistency, we follow the terminology of AIF360 to call one group privileged and the other unprivileged.

Which metrics to use?

To perform the bias analysis measurement, it is necessary to define which metrics are the most appropriate for the specific project.

In this project the analysis is going to focus on both direct and indirect bias with the goal of achieving group fairness. This decision has been taken, since direct bias could be present in some of the features (for example leeftijd and aantal_mannen), and there is always the risk of indirect bias.

The metrics considered were those that cover both disparate treatment and disparate impact, for example by averaging more specific metrics. Metrics designed to quantify only disparate treatment or only disparate impact were excluded. To select the best one, the following *Fairness Tree* (as suggested in the AIF360 documentation) has been consulted:



The model needs to be fair based on disparate errors, because an error means that a person would be investigated earlier than another. This also means that our interventions are punitive, and since all the cases are investigated independently of the outcome of the model we are intervening on the entire population.

These considerations drove us to the **false-positive rate ratio** as the main metric for this bias analysis. To understand how the false-positive rate ratio works, it is necessary to understand how the false-positive rate is calculated.

The **false-positive rate** is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events (regardless of classification):

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

It's the probability that a false alarm will be raised: that a positive result will be given when the true value is negative.

The false-positive rate ratio is defined as:

$$FPRR = \frac{FPR_{unprivileged}}{FPR_{privileged}}$$

The FPRR results can be analyzed in the following way: a FPRR

- between 0.8 and 1.25 means no bias;
- lower than 0.8 means that the privileged group is being discriminated;
- higher than 1.25 means that the unprivileged group is being discriminated.

* 0.8 and 1.25 have been chosen following the fourth-fifths rule.

How to do the analysis?

The best way to proceed is in two steps:

1. Analyze the bias of the features that carry direct bias.

   For example, the features *aantal_mannen* and *ratio_mannen* are a direct bias, since they map the category sex. These features can be analyzed by themselves and we can act on the bias that they contain.

2. Obtain sensitive data to check the indirect bias.

   The analysis of features indirectly biased is expected to be more accurate if instead of checking the bias in the features themself, the check is performed in the categories that originate the bias.

   For example, when analyzing a feature that maps nationality, we could either analyze the bias in this feature itself, or analyze the bias in the underlying characteristic that we're interested in, i.e. nationality. By doing the latter, obtaining a non-biased result allows concluding that none of the features used by the model are proxies for nationality bias.

Alpha algorithm Bias Analysis

Preliminary analysis

The data has been split into two parts: a *training set* and a *test set.* The reason is that, during the training phase, the model learns how to classify the data using the training dataset. After the model is trained, its success is measured using the test set.

There are multiple ways of performing this split, in this project two have been used:

- ***Temporal***: the data is ordered chronologically before performing the split, obtaining a train set that contains cases that happened earlier than any of the cases in the test set.
- ***Shuffle***: the data is shuffled randomly before the split, obtaining a mix of older and newer cases in both the train and the test set.

When the bias analysis started, the model was trained using a temporal split, and 20% of the data was left as the test set. The results were surprisingly biased in almost all the investigated categories, with an average false positive rate ratio distance of **62%** from the non-bias value. Such a negative result was suspicious, so a further investigation was needed.

The distribution of the values of all the features used by the model was analyzed, comparing the training and the test set. This analysis revealed that the feature 10.2.d contained zero values in the training set, but not in the test set. This means that all the cases used in the test set were opened when 10.2.d, while in the past, some cases were opened without a 10.2.d

To confirm the intuition that this was the reason why the model performed so biased in every aspect, the data was split using the shuffle technique. This is because the shuffle technique, mixing old and new cases, will mix in the test set cases with and without 10.2.d

A technical person could object that this could lead to fake better performances, because the model would learn information from the future and use that to classify a case of the past. This concern was addressed by checking that the model had the same performance using both splitting techniques, which was indeed the case.

The bias analysis of the model trained over the data split using the shuffle technique performed noticeably better, reducing the average false positive rate ratio distance to **21%**. This result is a big hint that the intuition was correct.

To double-check, the cases where 10.2.d was equal to zero were removed from the shuffled data. This check is to confirm that it was not the shuffling that reduced the bias, but the presence of cases without any 10.2.d in the dataset. The result was that the model performed even worse than using the temporal technique, achieving an average distance of **68%**.

From these analyses, it is possible to conclude that the processes used in the past to generate new cases, not relying only on the 10.2.d were important to reduce the bias of the investigations. For this reason, it is strongly recommended to restart those processes, since the creation of those extra cases reduces the bias.

For the analysis of the features, the shuffle split test set is going to be used.

Direct bias analysis

Sex

The analysis over sex has been done by analyzing the features *aantal_mannen* and *ratio_mannen.* This can be done, because both features involve counting the number of men living in the investigated apartments.

The split has been done setting as the *privileged group* if no women were living at the address, and the *unprivileged* if at least 1 woman was living at the address.

The analysis resulted in discrimination against the privileged group. So, instead of discrimination against women, in this specific case men were discriminated against. This means that the error committed by the model over cases where men were living in the apartment was bigger than the error over cases where also women were living in the apartment.

The outcome of this analysis combined with the low feature importance led to the removal of the feature itself. After removing the feature from the training of the model, the bias was checked again and it was successfully mitigated.

10.2.d



The result was the following:

Feature	False positive rate ratio
10.2.d [redacted]	1.11

Since the result is between 0.80 and 1.25, is possible to conclude that the model has no bias against 10.2.d Multiple other split values have been tried and they all resulted in a non-bias score.

Indirect bias analysis

Analyzing only the features used by the model is not enough to ensure discarding any kind of bias from our model. To increase the confidence of the result, data from BRP has been retrieved and used to analyze the bias of the model (not for training).

10.2.d

10.2.d

Feature	False positive rate ratio
10.2.d	1.10

Since the result is between 0.80 and 1.25, is possible to conclude that the model has an acceptable level of bias in 10.2.d

10.2.d

10.2.d



Feature	False positive rate ratio
10.2.d	1.77
10.2.d	0.23

The bias is introduced by the cases created because a 10.2.d , as explained before starting the analysis, and as expected this bias is reflected also in the feature itself.

To be sure that this is not a bias proxy of any of the relevant bias categories, a bias analysis over the extra features that cannot be used to train the model have been performed. The results can be seen in the *Indirect bias analysis* section.

10.2.d [redacted]

10.2.d [redacted]

Feature	False positive rate ratio
10.2.d [redacted]	1.17

Since the result is between 0.80 and 1.25, is possible to conclude that the model has an acceptable level of bias.

10.2.d [redacted]

10.2.d [redacted]

Here can be seen the list of the new characteristics analyzed and the results:

Feature	False positive rate ratio
10.2.d [redacted]	0.95
10.2.d [redacted]	0.90
10.2.d [redacted]	0.90
10.2.d [redacted]	0.88
10.2.d [redacted]	0.90
10.2.d [redacted]	0.90
10.2.d [redacted]	0.97
10.2.d [redacted]	0.95

Since the results are between 0.80 and 1.25, it is possible to conclude that the model has an acceptable level of bias for the new characteristics.

10.2.d [redacted]

10.2.d [redacted]

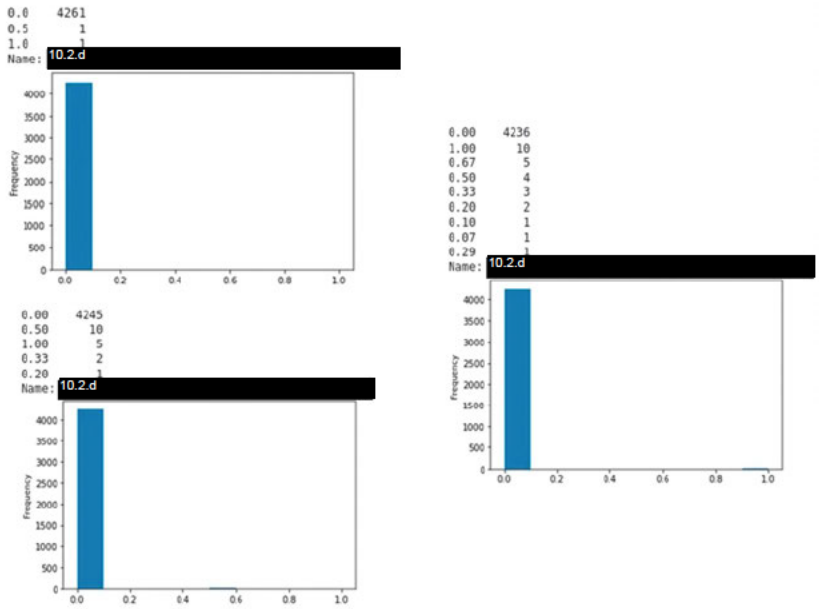
This is the result of the analysis:

Feature	False positive rate ratio
10.2.d [redacted]	*0.90*
10.2.d [redacted]	*1.13*
10.2.d [redacted]	*0.86*
10.2.d [redacted]	*0.90*
10.2.d [redacted]	*-*
10.2.d [redacted]	*0.95*
10.2.d [redacted]	*1.72*
10.2.d [redacted]	*1.34*

As can be seen, there are 3 features that deserve more investigation:

- 10.2.d [redacted]
- 10.2.d [redacted]
- 10.2.d [redacted]

These features were not scored at all or scored high due to the distribution of the values:



As it can be seen around the 99% of the values are equal to zero, so the bias analysis cannot perform its analysis correctly since the test data is not representative.

For the remaining features the bias is at an acceptable level.

Non-measurable categories

The following categories were not measured since the data was not available or collected:

1. Social class: related to the income of the inhabitants.
2. Religion.
3. Skin color.
4. Ethnicity.
5. Sexual orientation.
6. Political view.
7. Pregnancy.
8. Health.
9. Genetics.
10. Gender.
11. Disability.

Conclusion

Direct and indirect bias have been analyzed, obtaining useful insights regarding the model and the data used by it.

The first important outcome is a confirmation of the intuition that it's important to create cases using different processes, not only relying on reports from the citizens. These extra cases are necessary to reduce the bias of the investigations, and this has been proved by the analysis.

The current model has been trained on data containing both types of cases, therefore the error committed by the model is on an acceptable bias level. If in the future, the creation of cases through different methodologies other than the melding would stop, the dataset will be biased; and the model will probably learn the bias from it.

Met opmerkingen [10.2.e]: become instead of be?

The second important outcome is that men were discriminated against when the model was trained using sex as a feature. The error committed by the model over cases where men were living in the apartment was bigger than the error when also women were living in the apartment. The outcome of this analysis combined with the low feature importance led to the removal of the feature sex.

No other significant bias has been identified.

Considerations for the future

Bias in model outcomes vs. underlying data

The purpose of our bias analysis was to ensure that our holiday rental model is not biased. For that reason, we focused on detecting bias in the model outcomes and not in the underlying data.

Since we did obtain some results suggesting bias in the underlying data, it would be interesting in the future to analyze the cases opened as a result of citizens' reports to understand better the bias that those reports carry.

Investigating unavailable sensitive attributes

It would also be interesting to investigate what can be done to better analyze the bias on sensitive attributes that are not available within the municipality.

We did take the missing attributes into account by including them when making hypotheses of correlations that lead to indirect bias, but this is obviously not ideal. For instance, we could look for information about the missing attributes in other places, such as at the CBS, or see if we can approximate them by measuring other, correlated variables.

Creating groups

The analysis is furthermore limited by the choices that were made for the division of cases into privileged/unprivileged groups. For instance, we cannot be sure that no bias would have been found if the groups had been created differently.

We deem this risk small, since the groups were carefully crafted based on theory and the distributions seen in the data. Moreover, multiple variants were tested for the features that did not have clear splits.

Conclusie

De directe en indirecte bias zijn geanalyseerd. Hieruit zijn nuttige inzichten opgedaan over het model en de data die eraan ten grondslag ligt.

Het eerste belangrijke resultaat is een bevestiging van de intuïtie dat het belangrijk is om zaken te genereren via andere processen dan alleen de meldingen van burgers. Deze extra zaken zijn belangrijk om de bias van de onderzoeken te verminderen, zoals de analyse heeft bewezen.

Het huidige model is getraind op data die beide soorten zaken bevat, waardoor de fouten die het model maakt op een acceptabel niveau van bias liggen. Als in de toekomst het genereren van zaken door andere methoden dan de meldingen wordt gestopt, dan zal de dataset bias bevatten en het model zal deze bias uit de data leren.

Het tweede belangrijke resultaat is dat mannen werden gediscrimineerd wanneer het model werd getraind met geslacht als feature. De fout die het model maakte was groter voor zaken waarbij er alleen mannen in het appartement woonden dan wanneer er ook vrouwen woonden. Dit resultaat van de analyse gecombineerd met de lage belangrijkheid van deze feature heeft ertoe geleid dat de feature geslacht is verwijderd.

Er zijn geen andere substantiële biases gevonden.

Overwegingen voor de toekomst

Bias in uitkomsten van model vs. bias in onderliggende data

Het doel van de bias analyse was ervoor te zorgen dat het vakantieverhuur model geen ongerechtvaardigde bias bevat. Om die reden hebben we ons gericht op het detecteren van bias in de uitkomsten van het model en niet in de onderliggende data.

Omdat we desondanks een aantal indicaties hebben gevonden voor bias in de onderliggende data, zou het interessant zijn om in de toekomst de zaken nader te onderzoeken die zijn geopend naar aanleiding van een melding van een burger. Zo kunnen we de bias in deze meldingen beter begrijpen.

Niet-beschikbare gevoelige kenmerken onderzoeken

Het zou ook interessant zijn om te onderzoeken wat er gedaan kan worden om de bias te onderzoeken op gevoelige kenmerken die niet beschikbaar zijn binnen de gemeente.

We hebben rekening gehouden met deze niet-beschikbare kenmerken in de analyse door ze mee te nemen bij het opstellen van de hypotheses over correlaties die leiden tot indirecte bias, maar dit is duidelijk niet ideaal. We zouden bijvoorbeeld op andere plekken dan de gemeente naar informatie over de niet-beschikbare kenmerken kunnen zoeken, zoals bij het CBS, of kijken of we ze bij benadering kunnen schatten door andere, gecorreleerde variabelen te meten.

Groepen indelen

De analyse is ook ingeperkt door de keuzes die zijn gemaakt bij het indelen van zaken in gepriviligeerde en niet-gepriviligeerde groepen. We weten bijvoorbeeld niet zeker dat er geen bias zou zijn gevonden als de groepen anders waren ingedeeld.

We achten dit risico klein, omdat de groepen zorgvuldig zijn ingedeeld op basis van theorie en de verdelingen in de data. Daarnaast zijn meerdere varianten getest voor de features zonder duidelijke scheidslijn om groepen op te baseren.