Algorithm Fairness Exam: Communities and Crime

Johanne Rønby Sommer, Nina Sand Horup & Viktor Torp Thomsen
jors@itu.dk, nsho@itu.dk & vikt@itu.dk

Course code: KSALFAE1KU

2022-05-20

Contents

1 Introduction **1**

2 Data **1**
2.1 Sensitive and protected features 1
2.2 Creating binary features 1

3 Bias analysis **2**
3.1 Representation bias 2
3.2 High-crime neighbourhoods 2
3.3 Correlations in data 3

4 Methods **3**
4.1 Fairness metric 3
4.2 Model choice and data reprojection 4

5 Results **5**
5.1 Feature importances 5
5.2 Fairness and accuracy 6

6 Discussion **8**
6.1 Limitations of data debiasing method 8
6.2 Trade-off between fairness and accuracy 8
6.3 Violent crimes as a proxy 9
6.4 GOFAI and correlation machines 9
6.5 Possible racial difference in the ground truth 9

7 Conclusion **10**

1 Introduction

When we are making decisions in society, we are increasingly relying on big data and artificial intelligence to be part of the process. Examples of domains where we are already using big data are education, employment, advertising, health care, and policing, and the decisions which are being made have a real life impact on many people [8]. Therefore, it is vital that we understand how to implement fair algorithms, since it is by no means given that algorithms are fair by nature. In fact, it has already been shown several times, that some of the algorithms which are already implemented in society, are discriminating either groups of citizens or individuals [6, 13, 12, 9], even though it is by law prohibited to do so [1].

In this report, we examine how algorithms are prone to discriminating ethnic groups, if they are not properly monitored and if no interventions are made. In particular, we create an algorithm designed to predict crime rates in different neighbourhoods, and try to implement different initiatives to keep the algorithm from discriminating communities with a high ratio of African Americans, which it is otherwise prone to do if left to its own devices.

Link to Github where the data (and results) can be found: `https://github.com/ViktorTorp/afbe_project`.

2 Data

The Communities & Crime dataset combines socio-economic data from the 1990 US Census with law enforcement data from the 1990 US LEMAS survey, as well as crime data from the 1995 FBI UCR. It contains data from 1994 communities within 46 states, and is available at the UC Irvine Data Repository [2]. The value which is to be predicted in this data is the number of violent crimes per one hundred thousand citizens (a violent crime is defined as murder, rape, robbery, or assault). The data contains 127 columns besides the *ViolentCrimesPerCap* column. Of these, 4 attributes are categorical and 123 are numerical. The full list of data attributes can be seen in table 14 and 15. Note that the tables are taken from [14].

We remove the 23 features in the data which have missing values. We also remove the categorical features, since we find that they are not important in terms of predicting the crime rate. After dropping these columns, we are left with 95 feature columns, five protected columns and two target columns (since we both have the numerical- and the binary version of the crime rate).

2.1 Sensitive and protected features

In this project, we consider the percentage of African American people living in a neighbourhood to be the protected feature, but we have to account for several other features, because the protected feature is leaking into multiple other features: percentage of African Americans, percentage of Caucasians, percentage of Asians and percentage of Hispanics. Note that these columns do not necessarily add to one, since one individual can fit into several categories. However, we can generally infer the value of the percentage of African American citizens if we know the value of the other three columns. Therefore we consider all columns to be 'protected', even though we are focusing on the African American citizens.

It is important to note that the dataset contains additional features other than the percentage of African Americans, which are sensitive as well and which could therefore also be candidates for protected columns. Examples are the number of immigrants in the communities as well as what percentage of the police force that falls into the different racial categories.

2.2 Creating binary features

When we visualize the data as well as do some of the analysis, we create new binary columns based on the numerical columns *ViolentCrimesPerCap* and *racePctBlack*. We set a threshold for each,

[1] See for example the American anti-discrimination law

[2] `http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`

in order to fit each neighbourhood into the categories *HighCrime* and *IsBlack*. In other literature [11], they set the threshold for high crime to be the 70th percentile, and therefore we choose to do the same. This leads to a ratio of low-crime vs. high-crime neighbourhoods of 1411:583, which corresponds to an imbalance ratio of 2.42.

The threshold for *IsBlack* has been set to 0.06 [10] in other literature, but with no good explanation. We instead choose to set it so 0.15, since according to census data [3], between 13.4% and 16.2% percent of the American population is African American (depending on whether mixed races are taken into account).

3 Bias analysis

3.1 Representation bias

When we look at the data, we can see that the African American citizens are vastly underrepresented, since the mean percentage of African Americans in neighbourhoods is 18% (see fig. 1). If we look at the distribution of neighbourhoods categorized as predominantly non-African American vs. African American (the feature *IsBlack*), then the ratio is 1363:631, yielding an imbalance ratio of 2.16. However, when we compare these values to the percentage of citizens in the United States that are African American (see section 2.2), we see that this imbalance in the data actually originates from a population bias and not a selection bias. We actually have a slight over representation of African Americans in the data, but it is hard to say if this is statistically significant or not, since we do not know what percentage of all the neighbourhoods in the country are represented in the data.

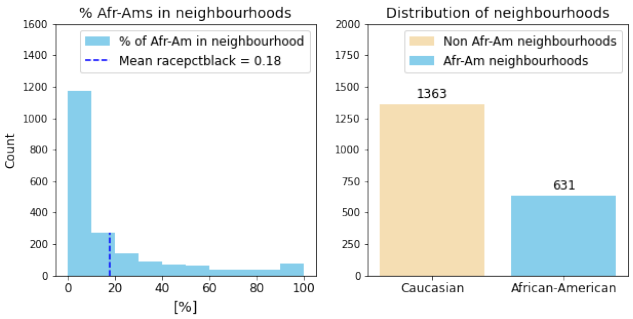


Figure 1: Graphs showing the under representation of African Americans in our data. We both look at the raw data in the percentages of African Americans in neighbourhoods (left) as well as the binary representation where we use the threshold of 0.15 to divide the data in two groups.

3.2 High-crime neighbourhoods

When we look at the distributions of the percentages of Caucasians and African Americans for both low and high-crime neighbourhoods, we can see that there is a major difference between the two population groups and how they relate to high-crime neighbourhoods, fig. 2. We see that the majority of the population in low-crime neighbourhoods tend to be non African American whereas the opposite is true when we look at high-crime neighbourhoods.

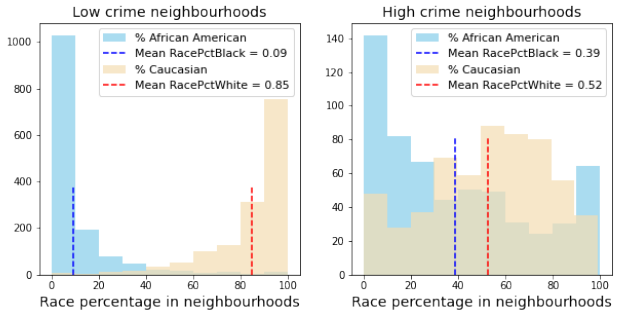


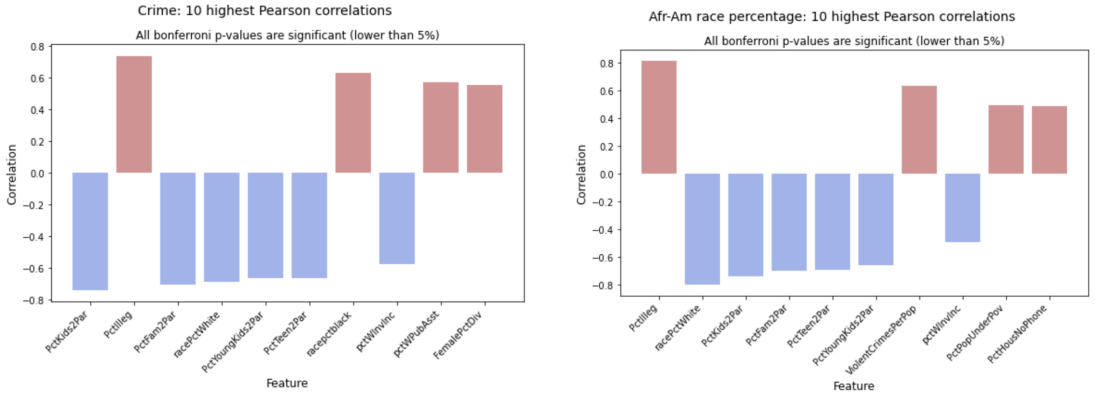
Figure 2: Crime distribution for the African American neighbourhoods compared to the non African American neighbourhoods. The distributions are almost opposite each other.

If we look at other features in the data, we can also see that the Caucasian and African American population groups are very diverse. Fig. 9 in the appendix shows how the African American population is the biggest group when it comes to poverty, unemployment, divorce, homelessness and lack of education. On the contrary, the Caucasians are leading when it comes to income.

[3] `https://www.census.gov/quickfacts/fact/table/US/PST045221`

3.3 Correlations in data

If we take a deeper look at why predominantly African American neighbourhoods are overrepresented in the crime statistics, we can see that there are a lot of features that mutually correlate with the crime rate as well as the percentage of African Americans in a neighbourhood.



(a) Top-10 Pearson correlations with crime.

(b) Top-10 Pearson correlations with the percentage of African Americans in neighbourhoods.

Figure 3: Correlation plots. Colouring refers to whether or not the correlation is positive or negative. It has been checked that all correlations displayed are statistically valid in the sense that their Bonferroni corrected p-values are below 5%

We can see that out of the top six features correlating with crime, five of the features are actually related to family situations and whether or not there are two parents in families. We see the exact same pattern when we look at correlations with the percentage of African Americans in the data. Surprisingly, *PctIlleg* (meaning how many percent of children are born with parents who are not married) correlates even more strongly with the percentage of African Americans than the percentage of Caucasians does, and this really underlines how much the family situations in general correlate with race.

4 Methods

4.1 Fairness metric

The choice of fairness metric to use for evaluation depends on which types of fairness we want to take into account.

First off, we must aim for the recall to be the same for African American neighbourhoods vs. other neighbourhoods. The recall measures how likely an area is to be predicted as having high crime, if it actually has high crime. This does not mean that we should actually police black and non-black neighborhoods equally much, only that the policing of areas with high crime should not be based on the the race of the people living there. It is also important to look at the false positive rate. The false positive rate checks how likely we are to predict high crime in areas which do not have high crime. This means that we are asking 'how likely are we to overestimate the crime rate in this area?'. Of course, this measure also has to be the same for black and non-black neighbourhoods, since this is exactly where the over-policing might be happening (when we are policing more in African American neighborhoods only because of the race of the people living there, and not because the crime rate is actually high). Since we choose to focus on ensuring that the recall and the FPR is the same for both groups, we use equalized odds to measure the overall fairness [4]. This means that our model must not only be equally good at identifying where there are high crime rates for the different groups, but it also cannot stigmatize one of the groups and raise more false alarms for that group compared to others. This ensures that the algorithm does not falsely accuse areas of having a high violent crime rate more- or less often based on the races of its inhabitants.

[4] https://developers.google.com/machine-learning/glossary/fairness$equalized_odds$

Another possibility could have been to look at the accuracy for both groups, and ensure that it was the same. However, if one group has a lot of crime in general, we can get a high accuracy just by always predicting that the crime rate is high, but the FPR would also be very high, and this would of course not be very fair.

4.2 Model choice and data reprojection

When we are using an algorithm to predict crime and thereby also possibly to determine whether to increase or decrease policing in an area, we are affecting people's day-to-day lives. Therefore, the ability to check why an algorithm has flagged one neighbourhood as having a high chance of having a lot of crime, is central. This is even underlined in the GDPR law which states: "When an individual is subject to 'a decision based solely on automated processing' that 'produces legal effects ... or similarly significantly affects him or her', the GDPR creates rights to 'meaningful information about the logic involved'" [2]. Therefore, we choose to use logistic regression as our primary model, because it relies on a simple linear combination of the input features, making it possible to examine the weights assigned to each feature. Hyper parameter tuning shows that an inverse regularization strength of 1 results in the highest accuracy. The model has been standardized before being given as input to the logistic regression model. The standardization is fitted to the train data and applied to both train and test data to avoid information leakage. In addition to the logistic regression model, we have used a decision tree. Simple tree based models can be visualized to see how each feature affects the predictions of the model. Whenever we evaluate the models' performance, we use a 5-folded stratified cross validating strategy[4, chapter 1.3].

To start off, we train the two models on the raw data to examine how they perform in terms of accuracy and fairness when we do not make any interventions. However, models often mirror the bias of the data they are trained on, and as we have shown already, our data contains a representation bias (caused by a population bias) as well as big structural differences between races caused by features which correlate highly with the target as well as the protected group. Therefore, a concern might be whether a model trained on the data without intervention will be biased against the protected group. Therefore, we explore how we can debias the data prior to training the models, in order to achieve fairer predictions. An initial naive approach to debiasing the data is to remove the protected features without any further modifications to the data. In our case, the protected features are all features indicating the percentage of a neighborhood being a specific race. However, this might not be sufficient since the models might derive the race features from other features, e.g. through correlations.

In [15] it is described how it is possible to reproject data in order to avoid data leakage between the protected features and the rest of the data due to correlations. The method is based on removal of correlations, and therefore it is only optimal for continuous data and not categorical data, but this is also what we have available in our data set. The approach is to represent all features as vectors and to reproject feature vectors onto vectors which are orthonormal to the protected feature vectors. This approach is represented in formula 1, where r_j denotes the j^{th} reprojected feature, x_j denotes the j^{th} initial feature before the reprojection, and p_i denotes the orthonormal vector to the i^{th} protected variable. In our case, we reproject our 95 features from the four protected features (the race variables).

$$r_j = x_j - \sum_{i}^{p} (x_j \cdot p_i)\, p_i \tag{1}$$

This reprojection will be strictly orthogonal to the protected feature but will also lower the models quality. To control the trade of between fairness and accuracy, a parameter λ was created such that a parameterized version of the reprojection is defined as

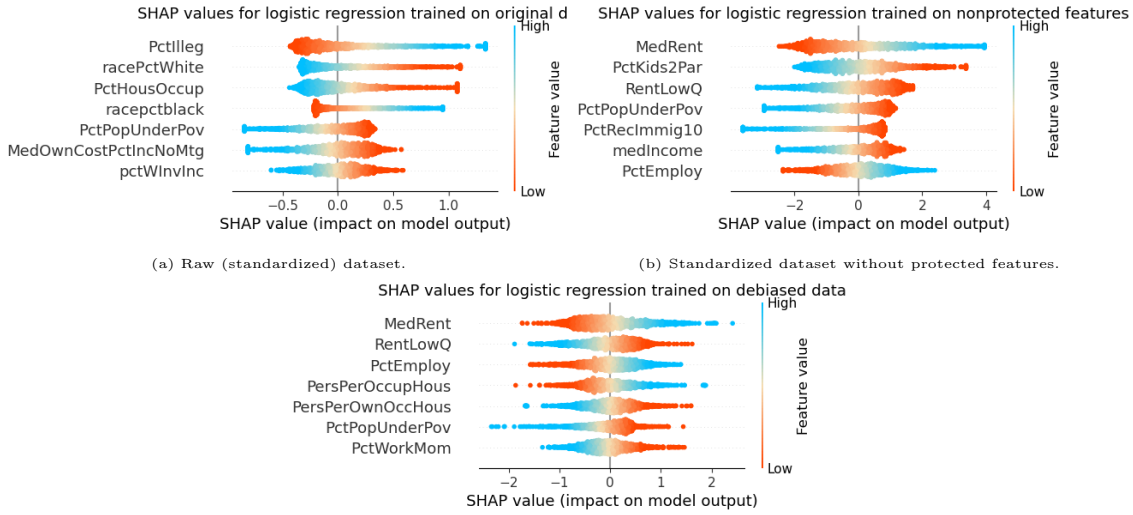
$$r\prime_j(\lambda) = r_j - \lambda(x_j - r_j) \tag{2}$$

Thus $\lambda = 0$ corresponds to $r\prime_j(\lambda) = r_j$ and $\lambda = 1$ is corresponds to $r\prime_j(\lambda) = x_j$.

5 Results

5.1 Feature importances

When looking at which features are deemed important by the logistic regression model trained on the original standardized data, it is evident that it is relying heavily on the protected features. If we look at the SHAP values for this model (see fig. 4a), we see that the feature with the highest impact on the model output is *racePctBlack* (percentage of American Americans in a neighbourhood). A high percentage of African Americans is the characteristic that pushes the prediction the furthest in the direction of the label 'high crime'. The percentage of African Americans has to be very low before it pushes the prediction towards 'False', and then it only has a limited impact on the model output. We can therefore conclude that using the raw data to train a logistic regression model results in the model making biased predictions.



(a) Raw (standardized) dataset.

(b) Standardized dataset without protected features.

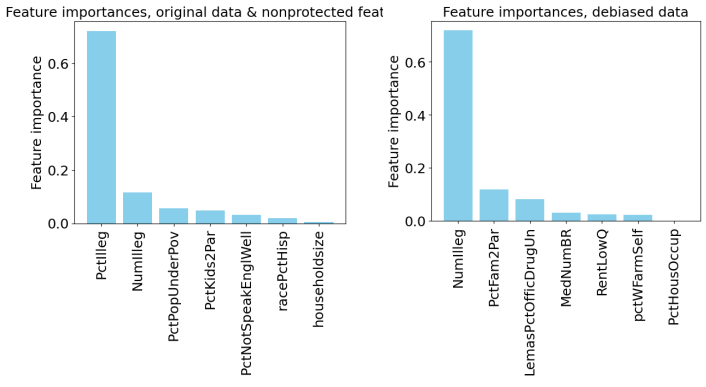
(c) Debiased data.

Figure 4: SHAP values for logistic regression model.

After removing the protected race features, the model can no longer use the former most impactful feature (measured in mean SHAP values), *racePctBlack*. Instead, *PctKids2Par* (percentage of kids in family housing with two parents) becomes the most impactful feature (see fig. 4b). The higher the percentage, the higher the probability of the prediction 'low crime'. From fig. 3a, we know that this feature is highly correlated with *racePctBlack*, meaning that it might just become a proxy for *racePctBlack*.

Removing the protected features can therefore not stand on its own if the goal is to create a fair data set to train a fair model, since there is a risk that the model derives race from other features that are highly correlated with race to begin with.

When training the model on the raw nonprotected features, *PctKids2Par* was the feature with the highest impact on the model output. The impact of the feature is slightly decreased after debiasing the data (see fig.



(a) Model trained on original data and on model trained on nonprotected features.

(b) Model trained on debiased data

Figure 5: Top 7 features with highest importance for decision tree model.

4c), but it is still the most important feature. This might indicate that the feature has some importance that cannot exclusively be attributed to the correlation with *racePctBlack*.

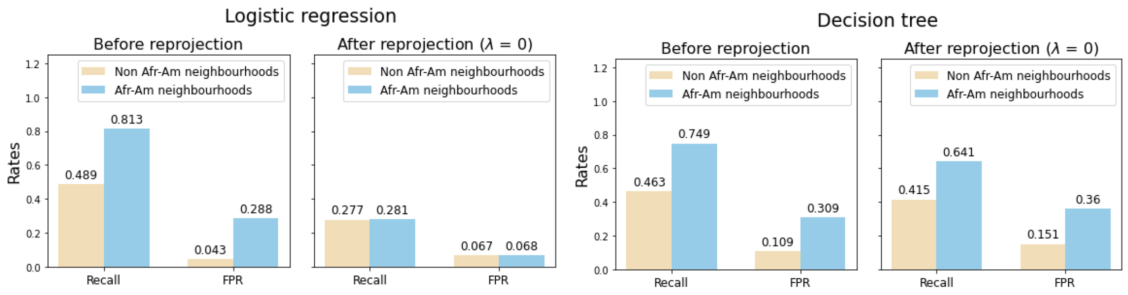
To ensure model interpretability, the depth of the decision tree is only 3, meaning that the number of features the model uses to make predictions is limited. The entire decision trees trained on respectively the original data and the debiased data can be seen in fig 10 and 11 in the appendix.

Fig. 5a shows the feature importance of the top 7 most important features for the decision tree model trained on the original data. The feature importance is measured as the decrease in node impurity weighted by the probability of reaching that node. The model primarily uses the feature *PctIlleg* (percentage of kids born to never married) to make decisions. *PctIlleg* is the feature with the highest Pearson correlation with *racePctBlack* (see fig. 3b), meaning that it might be used as a proxy for *racePctBlack*. Since the model does not use any of the protected features, removing the protected features does not change the feature importances. After debiasing the data, the model no longer uses *PctIlleg* to make decisions, but instead uses another version of it, *NumIlleg* (number of kids born to never married) (see fig. 5b). We will return to why this might be the case in the next section.

5.2 Fairness and accuracy

The reason behind debiasing the data was to make the models more fair. To check if we have succeeded, we can examine the equalized odds before and after debiasing the data. Fig. 6a shows the recall and the false positive rate (FPR), for the minority group *Predominately African-American (Afr-Am) neighbourhood* and the majority group *Predominately non African-American (Afr-Am) neighbourhood* for the logistic regression model. The left figure displays the results for the model which does not have access to the protected columns but where the data has not been reprojected yet. The right figure displays the results after the data was reprojected and with no protected features. The visualization shows that the model is significantly more unfair in terms of equalized odds before rereprojecting the data, since the recall and FPR is very different between the two protected groups. After reprojecting the features with $\lambda = 0$, we see that the model's recall and FPR are almost identical for both groups, thus having achieved more fair predictions when measured by equalized odds.

In fig. 6b, we show the performance of the decision tree model before and after reprojecting the data. Here we see that even though we have reprojected the data with $\lambda = 0$, we do not see any major improvement in terms of equalized odds.



(a) Equalized odds for the logistic regression model before and after data debiasing

(b) Equalized odds for the decision tree model before and after data debiasing

Figure 6

The improved fairness of the logistic regression model might come at the expense of lower accuracy. If we go back to looking at the SHAP values for the logistic regression model before and after debiasing the data (fig. 4a and 4c), there is a pattern where the SHAP values of the majority of the features are moved towards zero when debiasing the data, so that only a few samples have high impact on the predictions. When the impact of the majority of the features is being reduced, it makes it more difficult for the model to make predictions. This is being reflected in a decrease in accuracy for the logistic regression model when debiasing the data. Before reprojecting the data, the model had an accuracy of 0.72 for *Predominately (Afr-Am) neighbourhoods* and

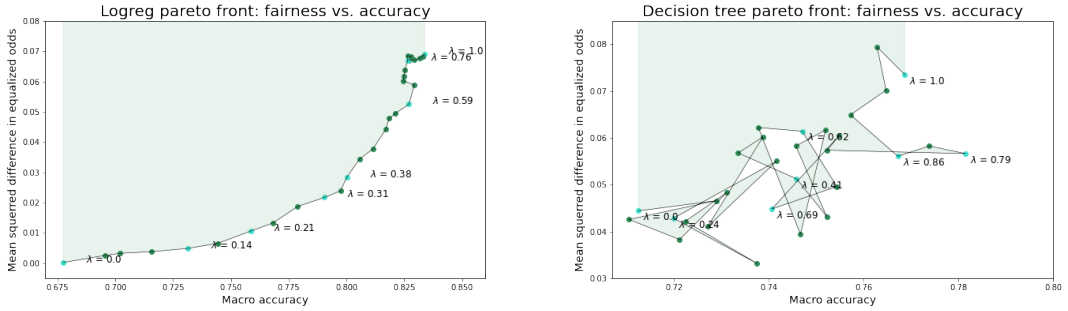
0.83 for *Predominately non African-American (Afr-Am) neighbourhoods*. After reprojection, the accuracies are 0.52 for *Predominately (Afr-Am) neighbourhoods* and 0.84 for *Predominately non African-American (Afr-Am) neighbourhoods*. We therefore see that the improvement in fairness, which is gained by reprojecting the data, negatively affects the model performance in terms of accuracy.

We can take a deeper look at this trade-off, if we evaluate the overall fairness as well as the overall accuracy for different values of λ. This is shown for the logistic regression model in fig. 7a as well as for the decision tree model in fig. 7b. For the logistic regression model, higher values of λ generally result in higher accuracy but also in higher unfairness. While the Pareto front for the logistic regression model is smooth, the relationship appear to be more random for the decision tree model.

The overall fairness is calculated as the mean squared error between the equalized odds for each group (formula 3), and the overall accuracy is calculated as the mean accuracy for each group, i.e. the macro accuracy (formula 4). G1 and G2 refers to each group respectively.

$$Unfairness_{overall} = \frac{\sigma_{recall} + \sigma_{FPR}}{2} = \frac{(G1_{recall} - G2_{recall})^2 + (G1_{FPR} - G2_{FPR})^2}{2} \tag{3}$$

$$Acc_{overall} = \frac{G1_{acc} + G2_{acc}}{2} \tag{4}$$



(a) Pareto front for the logistic regression.

(b) Pareto front for the decision tree.

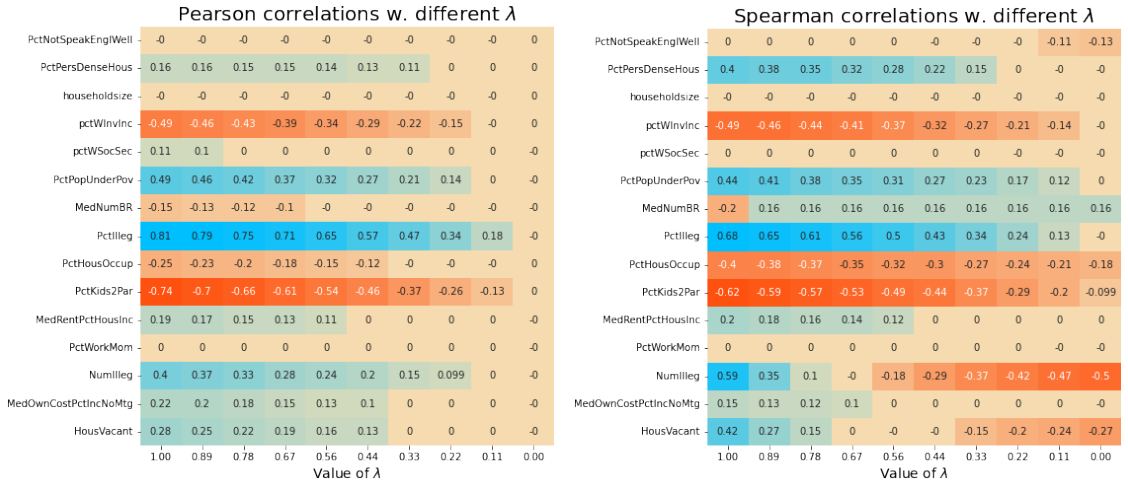
Figure 7: Pareto fronts displaying the balance between fairness and accuracy when training two different models on the debiased data using different λ-values. The shaded area indicates the area which is accessible to the model, whereas the unshaded area is the more optimal area where we have more fairness and more accuracy, but which we cannot achieve.

We believe that the reason why the decision tree does not obtain equalized odds like the logistic regression model does, and the reason why is it behaving so non-linearly when we look at the Pareto front, is that the decision tree model relies more on the Spearman correlations than on the Pearson correlations, since decision trees can make decisions and split the data based on the rank of the features.

The debiasing method only reduces the Pearson correlation between the protected and non-protected features, and not necessarily the Spearman correlation. Fig. 8a shows the Pearson correlations between the protected feature *racePctBlack* and other features as we debias the data with increasing values of λ. Fig. 8b shows the similar plot for the Spearman correlations. In these plots, we only show the correlations of a subset of features, consisting of the set of the 10 most important features from each model. We see that the features' Pearson correlations with *racePctBlack* move towards 0 as λ decreases, and that all of the features are completely decorrelated when $\lambda = 0$. The same relationship between λ and the correlations is not present when evaluating on the Spearman correlations. For several of the features, we see that even though decreasing λ results in the Spearman correlation initially moving towards 0, it continues to increase or decrease past 0 as λ moves towards 0. As a result, there is no value of λ where all features are completely decorrelated with the protected feature.

As mentioned, decision trees are sensitive towards Spearman correlations, and the fact that the Spearman correlations are not removed after debiasing the data, shows that the ranking of the data is not being completely changed. Instead, the debiasing only changes what features can be

used as proxies for the protected features by models that can exploit Spearman correlations. This is underlined by the fact that the feature with the most importance for the decision tree trained on the debiased data with $\lambda = 0$ is *NumIlleg* (see fig. 5b), which happens to be the feature with the highest Spearman correlation with *racePctBlack* (see figue 8b) when $\lambda = 0$. This shows that even though the Pearson correlations are removed, the decision tree might still be able to use features with high Spearman correlation as proxies. For the decision tree, changing the value of λ will therefore not have as significant an effect on the decision tree model as on the logistic regression model, and it will result in a more random relationship between λ and accuracy, and between λ and fairness for the decision tree model.



(a) Significant Pearson correlations for different values of λ.

(b) Significant Spearman correlations for different values of λ.

Figure 8: Significant correlations between the protected feature *racePctBlack* and the 15 most important features for both models, debiased with 10 different λ-values between 0 and 1.

6 Discussion

6.1 Limitations of data debiasing method

As shown in section 5.2, we found some limitations for the debiasing method, since it did not turn our to be a reliable way to obtain fair data for models which could exploit Spearman correlations. This limitation should be taken into consideration when doing model selection. For the given data set and debiasing method, the best choice of model will therefore be a logistic regression model, since the data debiasing is not enough to make a fair decision tree model, as can be seen from fig. 6b. Another limitation which should also be kept in mind, is the fact that Pearson decorrelation does not work for categorical values. This was not an issue in our case, since all of our data was categorical after the preprocessing, but it is very common to have many categorical features in a data set, and in these cases, this method has some additional limitations.

6.2 Trade-off between fairness and accuracy

As shown in section 5.2, it is possible to improve a logistic regression model's fairness for this particular dataset, by reprojecting the dataset as an attempt to remove its bias. However, as we see in fig. 7a, there is a trade-off between the model's accuracy and how fair its predictions are. By lowering λ, you gradually remove the Pearson correlation between the protected and non-protected features (as shown in fig. 8a), thus eliminating the model's ability to recognize and discriminate a particular group. This generally results in more fair predictions, when we evaluate the fairness with the equalized odds metric. But removing these correlations also lowers the model's capability of predicting the true target value and therefore simultaneously decreases the accuracy of the predictions.

The choice of λ is very important if the model is to be used to make decisions in real life. If λ is too low, the model might be very fair, but its predictive performance in terms of accuracy could

be so poor that the predictions cannot be trusted. However, if a too high value of λ is being used, the model might be very accurate but also unfair, which can lead to serious negative repercussion for certain minorities.

This leads us to a more fundamental discussion about fairness and about what we want to achieve with AI in society. In this report, we have always referred to fairness as treating two groups similarly, but we could also look at the fairness towards society as a whole. Having a low accuracy in terms of crime prediction might in the worst case scenario lead to underpolicing in some areas, and the consequences could be that people feel left to their own devices and do not feel protected by society. Determining how to balance the the fairness towards minorities versus the fairness towards all other citizens is extremely difficult. Even if we decided on some tangible philosophical solution such as Utilitarianism, where we would want to maximize the feeling of security across the board of all of society, how would we even measure how much some individuals gain vs how much others loose when we make certain decisions? All in all, the fundamental problem might be, that those who are in power and who are get to decide this balance, are usually from the majority and do not even consider this dilemma, and therefore society is missing out on having this discussion in the first place. We can see this problem, when we look at how tech companies are reluctant to having their algorithms audited and to responding to critique of their algorithms [6] and how half-hearted some attempts as fixing evidently unfair algorithms have been in the past [1]. Also, we are just now starting to realize and act on the unfairness of big data, even though it has been part of our society

6.3 Violent crimes as a proxy

It is important to give thought to the value we are predicting in this data set (the violent crime rate) and whether or not it would be suitable in different real life scenarios. If our algorithm would be deciding whether to increase or decrease policing in different areas (which is already an implementation we see in some police districts), then it might not be very suitable. In that scenario, we would be trying to predict the overall crime rate, and the violent crime rate might not be the best proxy in this case. This discrepancy as well as its possible consequences is important to remember when implementing algorithms in real life.

6.4 GOFAI and correlation machines

One of the first periods of excitement for artificial intelligence was caused by models which were created as expert systems using rule-based inference from knowledge-bases created by domain experts[5, C. 1]. In contrast to these early expert systems (often referred to as good old fashion AI, GOFAI) are the popular machine learning techniques of today. These popular models, such as logistic regression, decision tree, etc. are not created as logical encodings of domain knowledge, but rather as systems that learn to make decisions based on data. In other words, these models are developed to recognize patterns in data. Therefore, as shown in our experiments, if the data used to train these models contains e.g. racial biases, the models will be trained to continue the same racial inequalities. This poses a potential problem, even though a dataset might not contain sensitive features, structural inequalities present in the data can still be used as proxies for e.g. race or gender [7, P. 54]. As shown in section 5.2, the precautions of removing sensitive features are not enough to create a fair model if the data still contains proxies for the protected features.

6.5 Possible racial difference in the ground truth

A concern we had during this project, was if there was a racial component in the ground truth which could not be described by the correlations described in section 3.2. In other words; is there a partial correlation between race an crime rate which is not linked to any of the other features in the data? The reason we suspected this, is that there might be things such as unconscious bias or real-world feed-back loops at play (the more policing, the more crime is found), since this has been documented before [3]. On the other hand, we are only looking at very serious and violent crimes,

and it might be argued that a murder or an assault would be reported no matter how many police officers are present in an area already.

So, in order to determine if the difference in crime rate could be explained fully by our currently available features, or if there are additional influential factors, we compare neighbourhoods that are very similar to each other on all parameters except race and crime rate. Since we have a large amount of features, we use principal component analysis (PCA) to reduce the number of dimensions to two, in order to be able to do this comparison (see fig. 12a in the appendix). After transforming the data, we find 10 clusters with small euclidean distance in the PCA-space, containing one African American neighbourhood as well as five other neighbourhoods (see fig. 12b in the appendix). We choose to look at several non-African American neighbourhoods in each cluster in order to obtain information on the standard deviation within a cluster, i.e. how much the PCA similarity can be associated with similarity in terms of crime rate. When we analyse the ten clusters, we get the following results: The mean crime rate for all non African American neighbourhoods is 0.126, the mean standard deviation between the non African American neighbourhoods within the clusters is 0.076, and the mean crime rate for all African American neighbourhoods is 0.228.

This means, that when we compare African American neighbourhoods with other neighbourhoods which have similar feature representations, we still see a difference in crime rate (of approximately 0.1). This difference is bigger than the standard deviation we find if we compare five similar non-African American neighbourhoods with similar features. If the difference in crime rate between African American and non African American neighbourhoods had only been caused by the fact that the underlying features were also equally different, we would not have seen this difference. An example of the results within one cluster can be seen in fig. 13 in the appendix.

From this we can conclude that there is a measurable difference between African American and non African American neighbourhoods which is not linked to the features available in our data. There are two possible explanations; either we could be missing some latent features in the data set which can explain this difference, or there is something else which is causing this disparity (such as a real-world feed-back loop, the human unconscious bias or a third explanation). However, mapping out where this difference stems from is out of the scope of this project.

The question is, how we should deal with this discovery. We can conclude that the 'fairness' we are able to obtain by debiasing the data might not as fair as it looks, since its value is based on a possibly biased ground truth. A possible solution could have been to debias the actual crime with a sensible value for λ, which would rectify the imbalance, and then use this as the new ground truth used for calculating the fairness. However, we did not feel like that step should be taken before using more robust statistical methods for measuring the difference, and without knowing much about the origin of the disparity in the first place.

7 Conclusion

The goal of this paper was to create a fair algorithm for predicting the violent crime rate of neighborhoods given a range of socio-economic features describing the neighborhoods. We found that it was possible to create a logistic regression model that make predictions that do not discriminate on race by reprojecting the data before using it to train the model. However, the fairness comes at the cost of a decrease in accuracy. The trade-off between fairness and accuracy can be controlled in terms of how much to debias the data before training the model (see fig. 7a). The choice depends on the use case, since too much debiasing can lead to predictions with so low accuracy that they cannot be trusted, while too little debiasing can result in discrimination of minority groups and feedback-loops where the model reinforces negative stereotypes from the data.

The method we used to debias the data proved to have some limitations regarding model selection. It only removed Pearson correlations, which proved not to be enough for a decision tree model which is able to predict based on Spearman correlations. Another limitation to be aware of is that the ground truth crime rate of the training data of our algorithm might also be biased, making it impossible to make a 100% unbiased model. Finally, our model only predicts the violent crime rate, making it not suited for e.g. deciding where to increase or decrease policing, which could otherwise be an obvious use case for the model.

References

[1] Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. URL: `https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai`.

[2] Julia Powles Andrew D Selbst. Meaningful information and the right to explanation. *International Data Privacy Law*, 7:233–242.

[3] Sandra Bass. Policing space, policing race: Social control imperatives and police discretionary decisions. *Social justice*, pages 156–171, 2001.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[5] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014.

[6] A. Gilbertson & A. Sankin D. Mehrotra, S. Mattu, Dec 2021. URL: `https://themarkup.org/show-your-work/2021/12/02/how-we-determined-crime-prediction-software-disproportionately-targeted-low-income-black-and-`

[7] C. D'Ignazio and L.F. Klein. *Data Feminism*. Strong Ideas. MIT Press, 2020.

[8] Moritz Hardt, Sep 2014. URL: `https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de`.

[9] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner, May 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[10] Žliobaitė Calders Kamiran. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35:613–644, 2013.

[11] Roth Kearns, Neel and Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *Proceedings of Machine Learning Research*, 80:2564–2572, 2018.

[12] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA, 2016.

[13] Solon Barocas Andrew D. Selbst. Big data's disparate impact. *California Law Review*, page 671, 2016.

[14] V. Iosifidis & E. Ntoutsi T. L. Quy, A. Roy. A survey on datasets for fairness-aware machine learning. *CoRR*, abs/2110.00530, 2021.

[15] Kristina Lerman Yuzi He, Keith Burghardt. A geometric solution to fair representations. *AIES (Artificial Intelligence for the Earth Systems)*, 2020.

Appendix

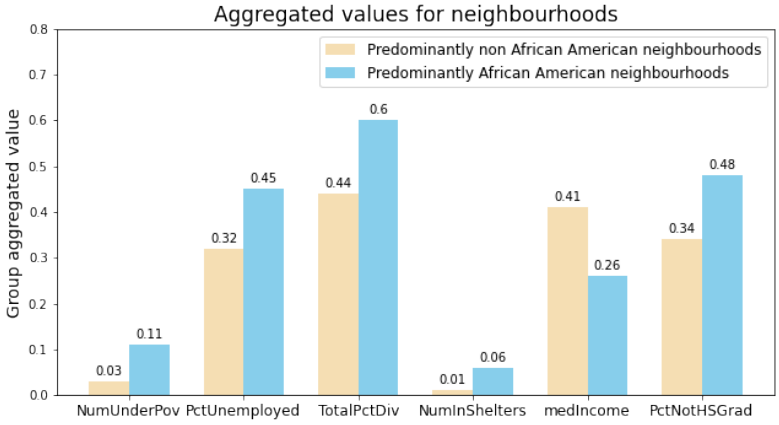


Figure 9: Different aggregated features and their value for Caucasians and African Americans respectively.

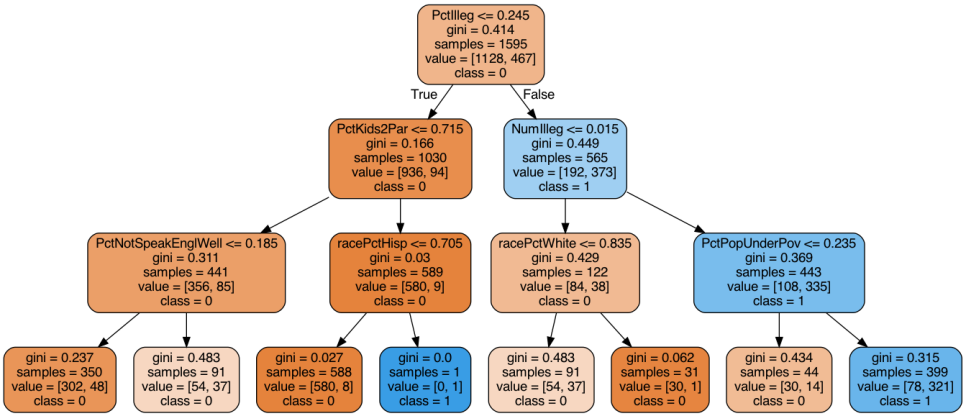


Figure 10: Visualization of decision tree trained on the original data. The colors indicate the majority class.

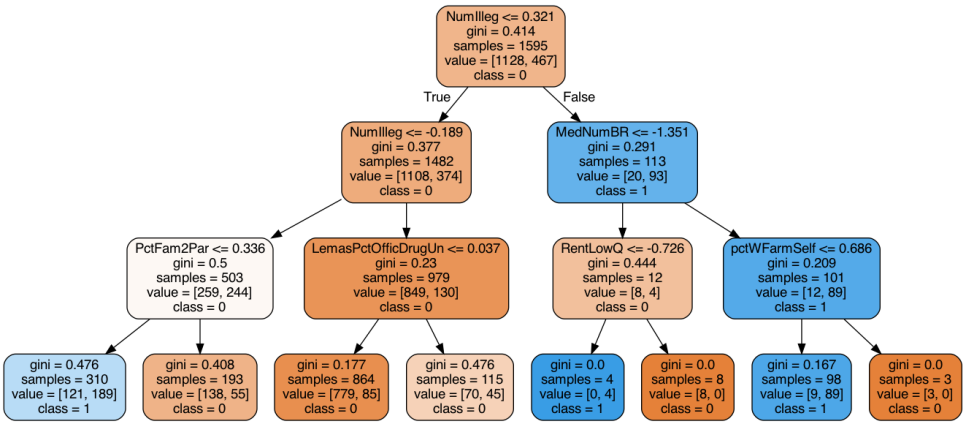
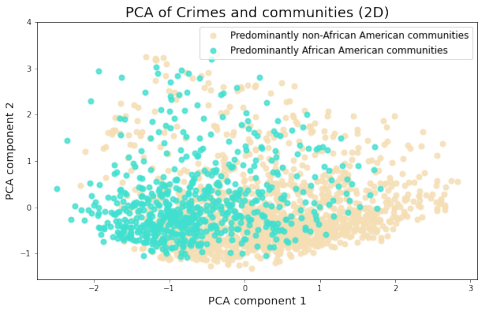
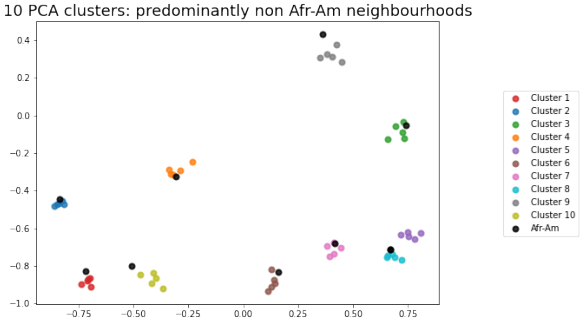


Figure 11: Visualization of decision tree trained on the debiased data. The colors indicate the majority class.



(a) 2D PCA transformation of the 85 features in the data (where protected columns have been excluded).



(b) 10 cluster of five non African American neighbourhoods and one African American neighbourhood with small euclidian PCA distance.

Figure 12: PCA transformation and clusters.

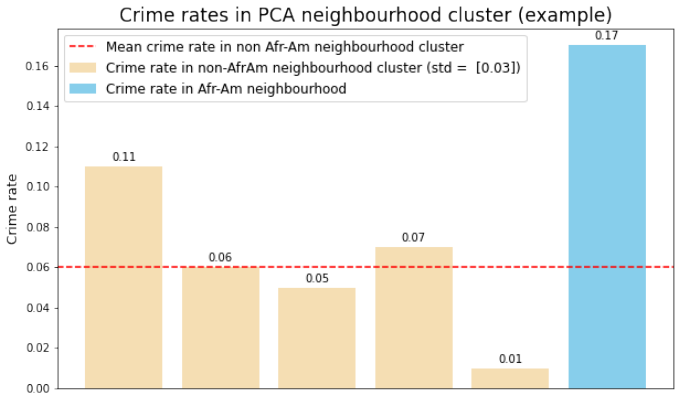


Figure 13: cc

Attributes	Type	Values	#Missing values	Description
racepctblack	Numerical	[0.0 - 1.0]	0	The percentage of population that is African American
pctWInvInc	Numerical	[0.0 - 1.0]	0	The percentage of households with investment/rent income in 1989
pctWPubAsst	Numerical	[0.0 - 1.0]	0	The percentage of households with public assistance income in 1989
NumUnderPov	Numerical	[0.0 - 1.0]	0	The number of people under the poverty level
PctPopUnderPov	Numerical	[0.0 - 1.0]	0	The percentage of people under the poverty level
PctUnemployed	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over, in the labor force, and unemployed
MalePctDivorce	Numerical	[0.0 - 1.0]	0	The percentage of males who are divorced
FemalePctDiv	Numerical	[0.0 - 1.0]	0	The percentage of females who are divorced
TotalPctDiv	Numerical	[0.0 - 1.0]	0	The percentage of population who are divorced
PersPerFam	Numerical	[0.0 - 1.0]	0	The mean number of people per family
PctKids2Par	Numerical	[0.0 - 1.0]	0	The percentage of kids in family housing with two parents
PctYoungKids2Par	Numerical	[0.0 - 1.0]	0	The percentage of kids 4 and under in two parent households
PctTeen2Par	Numerical	[0.0 - 1.0]	0	The percentage of kids age 12-17 in two parent households
NumIlleg	Numerical	[0.0 - 1.0]	0	The number of kids born to never married
PctIlleg	Numerical	[0.0 - 1.0]	0	The percentage of kids born to never married
PctPersOwnOccup	Numerical	[0.0 - 1.0]	0	The percentage of people in owner occupied households
HousVacant	Numerical	[0.0 - 1.0]	0	The number of vacant households
PctHousOwnOcc	Numerical	[0.0 - 1.0]	0	The percentage of households owner occupied
PctVacantBoarded	Numerical	[0.0 - 1.0]	0	The percentage of vacant housing that is boarded up
NumInShelters	Numerical	[0.0 - 1.0]	0	The number of people in homeless shelters
NumStreet	Numerical	[0.0 - 1.0]	0	The number of homeless people counted in the street
ViolentCrimesPerPop	Numerical	[0.0 - 1.0]	0	The total number of violent crimes per 100,000 popuation
state	Categorical	46	0	The US state (by number)
county	Categorical	109	1174	The numeric code for county
community	Categorical	800	1,177	The numeric code for community
communityname	Categorical	1,828	0	The community name
fold	Numerical	[1 - 10]	0	The fold number for non-random 10 fold cross validation
population	Numerical	[0.0 - 1.0]	0	The population for community
householdsize	Numerical	[0.0 - 1.0]	0	The mean people per household
racePctWhite	Numerical	[0.0 - 1.0]	0	The percentage of population that is Caucasian
racePctAsian	Numerical	[0.0 - 1.0]	0	The percentage of population that is of Asian heritage
racePctHisp	Numerical	[0.0 - 1.0]	0	The percentage of population that is of Hispanic heritage
agePct12t21	Numerical	[0.0 - 1.0]	0	The percentage of population that is 12-21 in age
agePct12t29	Numerical	[0.0 - 1.0]	0	The percentage of population that is 12-29 in age
agePct16t24	Numerical	[0.0 - 1.0]	0	The percentage of population that is 16-24 in age
agePct65up	Numerical	[0.0 - 1.0]	0	The percentage of population that is 65 and over in age
numbUrban	Numerical	[0.0 - 1.0]	0	The number of people living in areas classified as urban
pctUrban	Numerical	[0.0 - 1.0]	0	The percentage of people living in areas classified as urban
medIncome	Numerical	[0.0 - 1.0]	0	The median household income
pctWWage	Numerical	[0.0 - 1.0]	0	The percentage of households with wage or salary income in 1989
pctWFarmSelf	Numerical	[0.0 - 1.0]	0	The percentage of households with farm or self employment income in 1989
pctWSocSec	Numerical	[0.0 - 1.0]	0	The percentage of households with social security income in 1989
pctWRetire	Numerical	[0.0 - 1.0]	0	The percentage of households with retirement income in 1989
medFamInc	Numerical	[0.0 - 1.0]	0	The median family income
perCapInc	Numerical	[0.0 - 1.0]	0	Per capita income (national income divided by population size)
whitePerCap	Numerical	[0.0 - 1.0]	0	Per capita income for Caucasians
blackPerCap	Numerical	[0.0 - 1.0]	0	Per capita income for African Americans
indianPerCap	Numerical	[0.0 - 1.0]	0	Per capita income for native Americans
AsianPerCap	Numerical	[0.0 - 1.0]	0	Per capita income for people with Asian heritage
OtherPerCap	Numerical	[0.0 - 1.0]	1	Per capita income for people with 'other' heritage
HispPerCap	Numerical	[0.0 - 1.0]	0	Per capita income for people with Hispanic heritage
PctLess9thGrade	Numerical	[0.0 - 1.0]	0	The percentage of people 25 and over with less than a 9th grade education
PctNotHSGrad	Numerical	[0.0 - 1.0]	0	The percentage of people 25 and over that are not high school graduates
PctBSorMore	Numerical	[0.0 - 1.0]	0	The percentage of people 25 and over with a bachelors degree or higher education
PctUnemployed	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over, in the labor force, and unemployed
PctEmploy	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed
PctEmplManu	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed in manufacturing
PctEmplProfServ	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed in professional services
PctOccupManu	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed in manufacturing
PctOccupMgmtProf	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed in management
MalePctNevMarr	Numerical	[0.0 - 1.0]	0	The percentage of males who have never married
PersPerFam	Numerical	[0.0 - 1.0]	0	The mean number of people per family
PctWorkMomYoungKids	Numerical	[0.0 - 1.0]	0	The percentage of moms of kids 6 and under in labor force
PctWorkMom	Numerical	[0.0 - 1.0]	0	The percentage of moms of kids under 18 in labor force
NumImmig	Numerical	[0.0 - 1.0]	0	The total number of people known to be foreign born
PctImmigRecent	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigated within the last 3 years
PctImmigRec5	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigated within the last 5 years
PctImmigRec8	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigated within the last 8 years
PctImmigRec10	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigated within the last 10 years
PctRecentImmig	Numerical	[0.0 - 1.0]	0	The percentage of the population who have immigrated within the last 3 years
PctRecImmig5	Numerical	[0.0 - 1.0]	0	The percentage of the population who have immigrated within the last 5 years
PctRecImmig8	Numerical	[0.0 - 1.0]	0	The percentage of the population who have immigrated within the last 8 years
PctRecImmig10	Numerical	[0.0 - 1.0]	0	The percentage of the population who have immigrated within the last 10 years
PctSpeakEnglOnly	Numerical	[0.0 - 1.0]	0	The percentage of the population who speak only English
PctNotSpeakEnglWell	Numerical	[0.0 - 1.0]	0	The percentage of population who do not speak English well
PctLargHouseFam	Numerical	[0.0 - 1.0]	0	The percentage of family households that are large (6 or more)
PctLargHouseOccup	Numerical	[0.0 - 1.0]	0	The percentage of all occupied households that are large (6 or more people)
PersPerOccupHous	Numerical	[0.0 - 1.0]	0	The mean persons per household

Figure 14: Table showing the different features in the data (see next page for the rest of the features). The table is taken from [14]

Attributes	Type	Values	#Missing values	Description
PersPerOwnOccHous	Numerical	[0.0 - 1.0]	0	The mean persons per owner occupied household
PersPerRentOccHous	Numerical	[0.0 - 1.0]	0	The mean persons per rental household
PctPersDenseHous	Numerical	[0.0 - 1.0]	0	The percentage of persons in dense housing (more than 1 person per room)
PctHousLess3BR	Numerical	[0.0 - 1.0]	0	The percentage of housing units with less than 3 bedrooms
MedNumBR	Numerical	[0.0 - 1.0]	0	The median number of bedrooms
PctHousOccup	Numerical	[0.0 - 1.0]	0	The percentage of housing occupied
PctVacMore6Mos	Numerical	[0.0 - 1.0]	0	The percentage of vacant housing that has been vacant more than 6 months
MedYrHousBuilt	Numerical	[0.0 - 1.0]	0	The median year housing units built
PctHousNoPhone	Numerical	[0.0 - 1.0]	0	The percentage of occupied housing units without phone (in 1990)
PctWOFullPlumb	Numerical	[0.0 - 1.0]	0	The percentage of housing without complete plumbing facilities
OwnOccLowQuart	Numerical	[0.0 - 1.0]	0	Owner-occupied housing - lower quartile value
OwnOccMedVal	Numerical	[0.0 - 1.0]	0	Owner-occupied housing - median value
OwnOccHiQuart	Numerical	[0.0 - 1.0]	0	Owner-occupied housing - upper quartile value
RentLowQ	Numerical	[0.0 - 1.0]	0	Rental housing - lower quartile rent
RentMedian	Numerical	[0.0 - 1.0]	0	Rental housing - median rent
RentHighQ	Numerical	[0.0 - 1.0]	0	Rental housing - upper quartile rent
MedRent	Numerical	[0.0 - 1.0]	0	The median gross rent
MedRentPctHousInc	Numerical	[0.0 - 1.0]	0	The median gross rent as a percentage of household income
MedOwnCostPctInc	Numerical	[0.0 - 1.0]	0	The median owners cost (with a mortgage) as a percentage of household income
MedOwnCostPctIncNoMtg	Numerical	[0.0 - 1.0]	0	The median owners cost (without a mortgage) as a percentage of household income
PctForeignBorn	Numerical	[0.0 - 1.0]	0	The percentage of people foreign born
PctBornSameState	Numerical	[0.0 - 1.0]	0	The percentage of people born in the same state as currently living
PctSameHouse85	Numerical	[0.0 - 1.0]	0	The percentage of people living in the same house as in 1985 (5 years before)
PctSameCity85	Numerical	[0.0 - 1.0]	0	The percentage of people living in the same city as in 1985 (5 years before)
PctSameState85	Numerical	[0.0 - 1.0]	0	The percentage of people living in the same state as in 1985 (5 years before)
LemasSwornFT	Numerical	[0.0 - 1.0]	1,675	The number of sworn full-time police officers
LemasSwFTPerPop	Numerical	[0.0 - 1.0]	1,675	The number of sworn full-time police officers in field operations
LemasSwFTFieldOps	Numerical	[0.0 - 1.0]	1,675	The sworn full-time police officers in field operations per 100,000 population
LemasSwFTFieldPerPop	Numerical	[0.0 - 1.0]	1,675	The number of sworn full time police officers in field operations
LemasTotalReq	Numerical	[0.0 - 1.0]	1,675	The total requests for police
LemasTotReqPerPop	Numerical	[0.0 - 1.0]	1,675	The total requests for police per 100,000 popuation
PolicReqPerOffic	Numerical	[0.0 - 1.0]	1,675	The total requests for police per police officer
PolicPerPop	Numerical	[0.0 - 1.0]	1,675	The number of police officers per 100,000 population
RacialMatchCommPol	Numerical	[0.0 - 1.0]	1,675	A measure of the racial match between the community and the police force
PctPolicWhite	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are Caucasian
PctPolicBlack	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are African American
PctPolicHisp	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are Hispanic
PctPolicAsian	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are Asian
PctPolicMinor	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are minority of any kind
OfficAssgnDrugUnits	Numerical	[0.0 - 1.0]	1,675	The number of officers assigned to special drug units
NumKindsDrugsSeiz	Numerical	[0.0 - 1.0]	1,675	The number of different kinds of drugs seized
PolicAveOTWorked	Numerical	[0.0 - 1.0]	1,675	Police average overtime worked
LandArea	Numerical	[0.0 - 1.0]	0	Land area in square miles
PopDens	Numerical	[0.0 - 1.0]	0	The population density in persons per square mile
PctUsePubTrans	Numerical	[0.0 - 1.0]	0	The percentage of people using public transit for commuting
PolicCars	Numerical	[0.0 - 1.0]	1,675	The number of police cars
PolicOperBudg	Numerical	[0.0 - 1.0]	1,675	Police operating budget
LemasPctPolicOnPatr	Numerical	[0.0 - 1.0]	1,675	The percentage of sworn full-time police officers on patrol
LemasGangUnitDeploy	Numerical	[0.0 - 1.0]	1,675	Gang unit deployed
LemasPctOfficDrugUn	Numerical	[0.0 - 1.0]	0	The percentage of officers assigned to drug units
PolicBudgPerPop	Numerical	[0.0 - 1.0]	1,675	Police operating budget per population

Figure 15: Table showing the different features in the data (see previous page for the rest of the features). The table is taken from [14]