Milestone 4 Report

Álex Socías, Jean Devillard & Nathan Fiorellino

1 Abstract

The idea of the original paper was to evaluate a racial profiling trend across the United States. Our aim was to develop a model able to predict if a search will be conducted or not on a subject, since being able to make precise predictions based on external variables could indicate bias in search decisions. We developed the model on the Washington state, but this is mostly an example, and our general approach could be used on different states. Thus we performed a decision tree analysis of data on the Washington state. We then evaluated the precision of the model compared to other methods, how the model evolved over different time periods, and if it was generalizable to another state.

On all methods used, we noticed a precision cap on the testing set, and all trials to improve this precision have resulted in either no significant change or an overfit. This precision cap shows that a search decision has a huge human factor which we cannot properly account for, no matter the method used, but also ensures that our model is as precise as possible. The decision trees seem to confirm the search papers hypothesis of racial bias in search decisions, but also suggest that the subject age is the most important factor in predicting if a search will be conducted or not. The decision trees seem to be relatively stable over time.

2 Introduction

We tried to analyze global behavior patterns of the police officers during conducted searches and predict if a driver will be searched. Here we performed decision tree analysis. The idea behind the decision tree is to measure the influence of certain variables on a decision, which in this case is the search of the subject. The variable we chose to take in account are the subjects age, race and sex. In the same way as the paper we restricted ourselves to White, Black and Hispanic races. We first fitted the tree to different time intervals for stops of the Washington state patrol and then tried to contrast this tree to other states.

Before explaining our analysis and methods, the behavior of the police between states might differ because of specificities. We think that the model precision will be limited because of the human factor in search decisions, which we cannot account for. We believe that it is interesting to see if any of the factors or physical traits can induce the police to behave differently against a driver and deviate in a case of racial or sex disparity. Being able to precisely predict if a search will be conducted or not does not necessarily indicate bias, but rather corroborates this hypothesis.

3 Related work

A related study of the racial disparities is the paper on which we performed our extension. In the original paper police stop data from all the United States was evaluated. The results were that black drivers were less likely to be stopped after sunset, when a 'veil of darkness' masks one's race, suggesting bias in stop decisions. Furthermore, by examining the rate at which stopped drivers were searched and the likelihood that searches turned up contraband, they found evidence that the bar for searching black and Hispanic drivers was lower than that for searching white drivers. Finally, they found that legalization of recreational marijuana reduced the number of searches of white, black and Hispanic drivers, but the bar for searching black and Hispanic drivers was still lower than that for white drivers post-legalization. The overall results indicate that police stops and search decisions suffer from persistent racial bias and point to the value of policy interventions to mitigate these disparities.

We can related this result obtained with the actual situation in North America where demonstrations and protests by a large part of the United States population against racial disparity in police action have been carried out. Due to recent events such as the case of the death of George Floyd that has made that this necessary social revolution take more force in order to mitigate the disparities present in many other human activities that have less visibility a be able to live in a fairer world where people are treated in the same way regardless of their race or origin. We strongly believe that every little step we take to improve disparities and injustices brings us closer to living in harmony with all human beings.

4 Data collection

The data that was collected comes from the same page of the original dataset that the racial disparities paper use. Standardized stop data are available to download (by location) provided in both CSV and RDS formats.

In addition, shape files are available for select locations. We can distinguish different sources of the information collected by the police, data from state patrol or from municipal police. In our expansion, we used the data from the Washington state patrol. We were interested in the influence of the marijuana legalization status, so we chose the Washington state patrol because it is the most extensive statewide data on a legal state.

All data to reproduce the findings of this study are available at https://openpolicing. stanford.edu.

5 Data set & Summary Statistics

More than 7million of Americans are from the state of Washington only near of a 2 % of all the American population. The data with we are dealing in this extension is extensive with 30 different features for the state of Washington.

The first step with the raw data was creating a new dataframe with only our features of interest. These features are the year, the subject age, the subject race, the subject sex and if a search was conducted. Furthermore, we cleaned out all rows with missing values, and reduced the subject races to white, black and hispanic.

We then splitted our dataset and created new ones that contain information on two years each. We took out the first year and the last year to be sure that the data is as complete and homogeneous as possible. The year of marijuana legalization in the state of Washington was 2012 (December 6, 2012), and to be sure that the data collected is not affected, we do not consider the year of 2012 and 2013 in our analysis cause in this period like the original article says, they were changes of police behavior. With this fact we know that in different periods of data analyze perhaps we can isolate specific trends. The behavior of the police not only depends of the subject's personality, but their way of acting may be influenced by uncontrolled political or social factors that could influence our conclusions. Thus, we decided to separate the entire data set obtained on two year time spans in order to limit the influence of these factors.

6 Methods

In this extension, we performed a Decision Tree Analysis on the stops of the Washington state police, with the help of the sklearn library. Then compared the accuracy, precision and recall between the different models with the SGD classifier and logistic regression. We also performed an optimization based on validation set precision (Grid search cross validation) to get best hyperparameters to fit our model with the Decision Tree Analysis. We also reproduced our optimized tree on different states calculating missing data ratios if it was necessary.

Decision trees are a common tool to help identify a strategy most likely to reach a goal. In our case it is useful to see the behavior of the police. Compared to all different models, the decision tree is a satisfying possibility. It has the merit to be extremely graphical and simple to understand. It also can be used to draw conclusions by hand. Furthermore, associations between the different variables are automatically taken into account, which is not the case in logistic regression. Random forests are usually more accurate, but it is a black box model (not graphical at all, and impossible to use manually). We tested externally a random forest, and there was no increase in precision on the testing set no matter the hyper parameters chosen. Decision trees used in data mining are of two main types, regression tree and classification tree. In our case we used the classification tree where the analysis is when the predicted outcome is the class (discrete) to which the data belongs.

We used the grid search cross validation in order to find the best hyperparameters of criterion and splitter function of the function imported from sklearn (DecisionTreeClassifier). The criterion have the function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain. Furthermore, the splitter is the strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.

Others methods used in this evaluation were the Logistic Regression and the SGDClassifier from sklearn. Logistic Regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. More concretely it is a statistical model that in it's basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). The SGDClassifier is an estimator that implements regularized linear models with stochastic gradient descent (SGD) learning: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate). SGD allows minibatch (online/out-of-core) learning via the partial fit method. For best results using the default learning rate schedule, the data should have zero mean and unit variance.

We used these other methods because the search rate is extremely low, which means that the data is extremely unbalanced. Therefore, we evaluated the different models based on precision, since precision is the ratio of true positives on predicted positives (positive: search conducted) and as such is not affected by a large proportion of negatives. On all methods used, there seems to be a precision cap on the testing set. This precision cap is always the same, and all trials to improve this precision have resulted in either no significant change or an overfit. This precision cap can be interpreted as an impor-

tant human factor in the search decisions which cannot be properly accounted for in our models. The regularity of the precision for all models can be explained by the important number of samples. The precision in itself is average, but considering the small proportion of positives and the importance of the human factor it is actually satisfying.

7 Results and findings

Decision tree, SGDClassifier and logistic regression analysis on the dataset for the whole time period had a prediction accuracy of 0.970, precision of 0.485, and recall of 0.500 on both training and testing sets. This regularity can be explained by the important number of samples (more than 11 million entries before splitting). These results are extremely average, however they can be explained by both the very small search rate (approximately 3%) and the importance of the human factor in search decisions. There seems to be a precision cap, and all trials to improve this precision have resulted in either no significant change or an overfit. This precision cap can be interpreted as an important human factor in the search decisions which cannot be properly accounted for in our models. The fact that it is the same limit for each method used shows that our manipulations were correct. The regularity of the precision for all models can be explained by the important number of samples. The precision in itself is average, but considering the small proportion of positives and the importance of the human factor it is actually satisfying.

We found that the best function for these hyperparameters are "gini" in the case of the criteria and "best" in the case of the splitter. However, the grid search returned an optimal level of one. Therefore, it was not really conclusive. Previous trials where we took into account the legalization status for the whole time period returned a max depth of 10. Therefore, the main conclusion is that the depth of the decision tree has very little influence on the precision in our situation. Hence, we used 3 levels since we have 3 variables, which was an arbitrary choice but we deemed it sufficient to have an idea of the relationship between variables.

All 3 decision trees we generated were similar to the one shown in **Fig1.** The only difference are the thresholds for each root node, but since both race and sex were categorical data that we performed linear encoding on, the conclusions of the decision trees are not modified. This regularity over time means that the main trends isolated stay relevant, and also that marijuana legalization did not significantly modify these trends. We can notice that one of the main variables to influence the search decision is the subject age. This is intuitively consistent, since policemen are relatively unlikely to search an elderly person. Here, the older a person is, the less likely this person is to be searched. For young people, the major determining criterion seems to be if the person is Hispanic or black, then the sex of the person. This suggests the racial bias trend investigated by the paper is founded, but that there might also be a sexual bias to take into account in future studies. For middle aged persons, only race seems important, specifically if the person is black or not. However this is at 3 levels, so importance of the sex might just be hidden. Thus, there seems to be a bias against black or Hispanic drivers for young subjects, and a bias against black drivers for middle aged subjects. Additional statistical testing is necessary to assert the actual influence of these variables on a search decision, and the trends we observed do not necessarily mean that there is a racial bias in searches, as preferential searching of one race or sex over another could be due to external factors not accounted for here. Our results are in line with the findings of the paper, and the possibility of a sexual bias needs to be further investigated.

When used on the same time periods but with Colorado state patrol data, the different trees had a fit accuracy between 0.994 and 0.998, precision between 0.497 and 0.499, and recall of 0.500. This means that the different trees had a better fit on the Colorado data that on the Washington data, on which they were trained. This is a surprising result, but we can draw two main conclusions from it. The first one is that variations in the prediction quality such as the human factor are highly state specific. The second one is that thee main search decision trends seem to remain the same from one state to another. This also shows that our method is relatively general, and can be reused for a wide range of data.

8 Conclusion

The main conclusions that we extract for our analysis is that in the behavior of the police there is a human factor that is impossible to predict and quantify with a decision tree classifier. Also we can say that this fact occurs in the two states assessed (Washington and Colorado) and should be relevant along all the United States. Furthermore, we find bias against black and Hispanic people as the original paper but also the decision tree shows that the most important factor to a search decision is the subject's age, and that sex also plays a role in this decision. The trends isolated in our decision trees seemed stable over time, and not affected by marijuana legalization. Finally, there seems to be a high level of state specificity in these search trends. Although we cannot use this model to predict if a subject will be searched for sure, it is in line with the papers findings. Our general approach seems valid, and can be reused for different states. Additional research should be performed in order to evaluate the sexual bias in these searches.

9 Figures

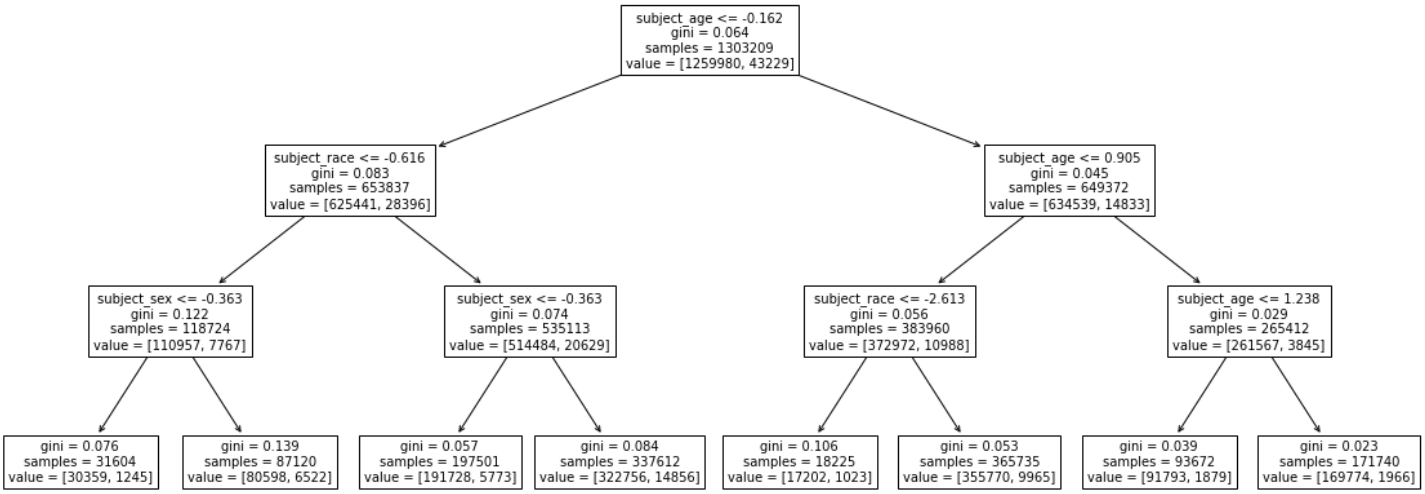

Figure 1: Decision tree for WA state patrol for 2010 and 2011

10 References

Pierson, E., Simoiu, C., Overgoor, J. et al. A large-scale analysis of racial disparities in police stops across the United States. Nat Hum Behav 4, 736–745 (2020).

https://doi.org/10.1038/s41562-020-0858-1

https://en.wikipedia.org/wiki/Decision_tree

https://en.wikipedia.org/wiki/Decision$_t ree_l earning$

https://en.wikipedia.org/wiki/Data_mining

Data : https://openpolicing.stanford.edu/data/