Introduction

In this report, we will describe a new model that helps police department conducting its stop and search policy of vehicles in the US. The police department has received lots of complaints about

its stop and search policy because, according to critics, their search policy is biased against people of certain backgrounds.
The following work proposes a study of these criticisms. We analyze if they are valid, and we suggest a method that approaches this problem of discrimination while also providing a good model for predicting if there is contraband inside a car.

Client requirements

Summary

The objectives given by the police department were:
"(1) determine whether these criticisms seem to be substantiated, and (2) create a service to fairly decide whether or not to search a car, based on objective data".
And the requirements were the following:

- A minimum 50% success rate for searches (when a car is searched, it should be at least 50% likely that contraband is found)
- No police sub-department should have a discrepancy bigger than 5% between the search success rate between protected classes (race, ethnicity, gender).
- The largest possible amount of contraband found, given the constraints above.

Requirements clarifications

(Business + TechnicalConsider that this is what you send the client ahead of time to close ambiguities from the busienss requirements. It should turn ambiguous business requirements into hard technical requirements.)
Regarding the initial requirements of the client, there are thre considerations in order to clarify our interpretation of the goals:

- "A minimum 50% success rate for searches" means a minimum precision of 50% TP/(TP+FP)
- "No police sub-department should have a discrepancy bigger than 5% between the search success rate between protected classes (race, ethnicity, gender)". We cannot guarantee this condition for all the departments. For example, departments with few observations for which we could not learn if there were real discrimination happening, will probabily not meet this condition, since it is not possible to correct it with the current amount of data.
- "The largest possible amount of contraband found" means the maximum possible recall: how many contraband we recalled from all cars with contraband.

Dataset analysis

General analysis

In this section we will give an overview of the dataset in order to build the foundation for the next sections.

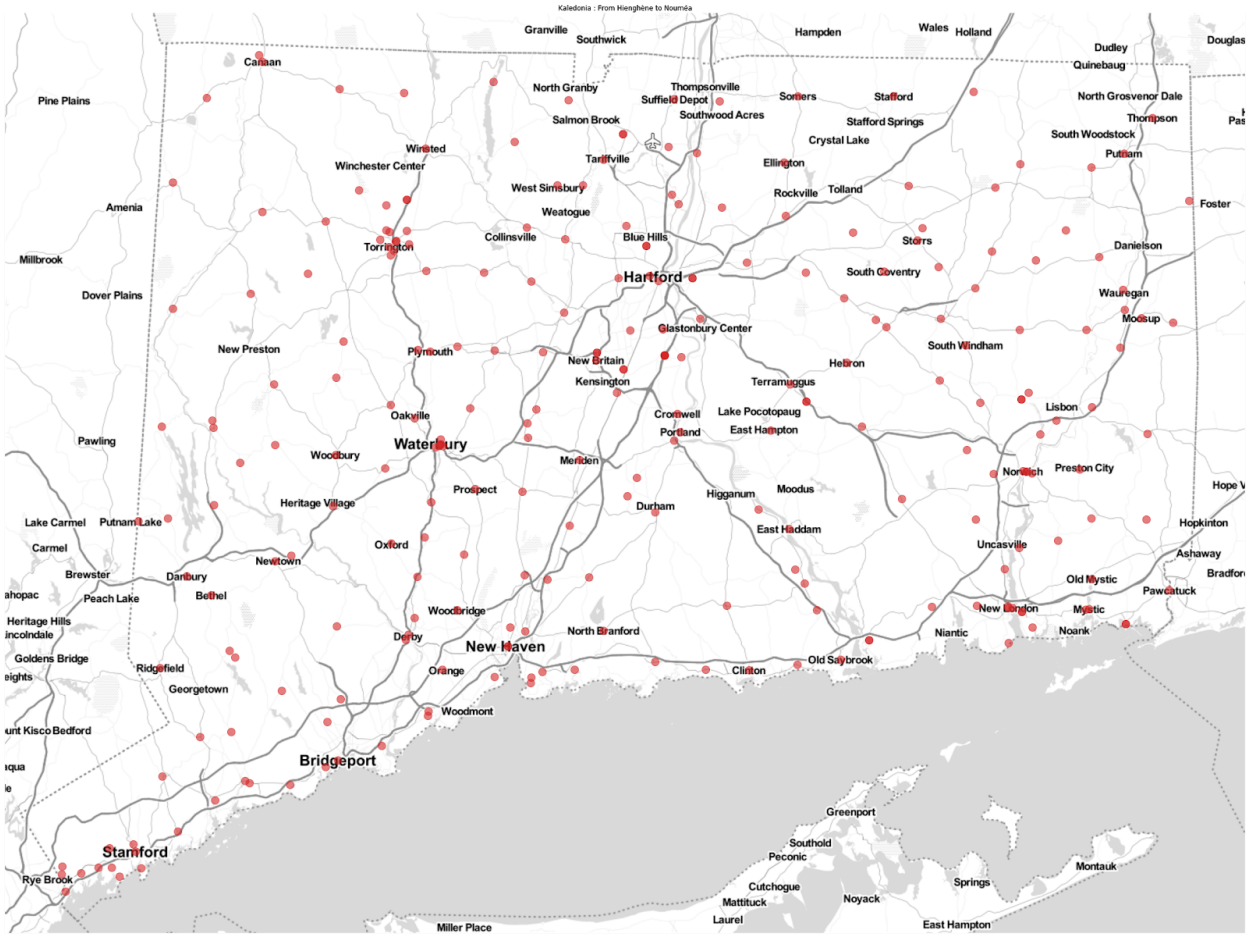
Description

These are the variables present in the dataset.

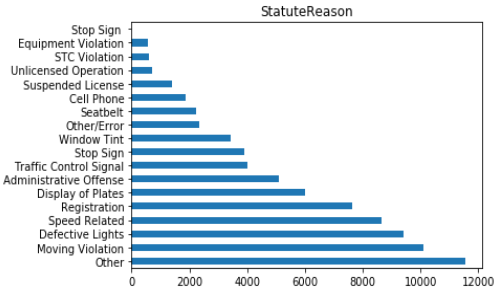
Variable name	Description	Codes used
VehicleSearchedIndicator	Whether the vehicle was searched	
ContrabandIndicator	Search Disposition: Contraband and/or evidence discovered	
Department Name	Name of the police department	
InterventionDateTime	Date and time of the intervention	
InterventionLocationName	Location of the intervention	
InterventionReasonCode	Code for the reason given for stopping the vehicle	Investigation: I Violation: V Equipment: E
ReportingOfficerIdentificationID	Reporting Officer Identification ID	
Resident Indicator	Whether the subject was a resident of the state	
search_authorization_code	Authority to search vehicle	N-Not Applicable C-Consent I-Inventory O-Other:Probable Cause, Reasonable Suspicion, Plain View Contraband, Incident to Arrest, Drug Dog Alert, Exigent Circumstances
StatuteReason	Reason given for stopping the car	
SubjectAge	Age of the main occupier of vehicle	
SubjectEthnicityCode	Officer perception of the ethnicity of subject	Hispanic: H Middle Eastern: M Not Applicable: N or N/A
SubjectRaceCode	Officer perception of the race of subject	W - White B - Black I - Indian America/Alaskan Native A-Asian/Pacific Islander U - Unkown
subject_sex_code	Subject Sex Code	
TownResidentIndicator	Whether the subject was a resident of the Town.	
InterventionLocationName	Intervention Location Name	

The dataset has 2.473.643 entries of stopped cars. However, only 76.743 of them were actually searched for contraband. These are the ones we are mainly interested in since, for the interest of creating a predictive system, we only know if the car had actually contraband in its interior after it is searched. Also, in our context, we will only receive observations if the police officer already wants to

search the car, so the populations should be similar. From now on, we will refer to these observations as the whole dataset because of this reason.
These data points extend in time, from 2013-10-01 to 2018-05-1 and after studying the locations (which had some ambiguous names that corresponded to multiple cities in the USA), we found out that all the unambiguous cities were from Connecticut, so we assumed that all the places corresponded to that state, and almost all of them existed in CT, USA.



Race and Ethnicity		Gender		Residency		Age	
White	47.9%	**Male**	81.4%	**Connecticut Resident**	40.2%	**<21**	19.9%
						22 to 30	41.6%
Black	28.2%					**31 to 40**	20.9%
Hispanic	23.0%	**Female**	18.6%	**Nonresident**	59.8%	**41 to 50**	10.2%
						51 to 60	5.6%
Other	0.9%					**60+**	1.8%



Business questions Analysis

The two objectives that we are trying to achieve ("(1) determine whether these criticisms seem to be substantiated, and (2) create a service to fairly decide whether or not to search a car, based on objective data") can be decomposed into small steps with different challenges each that we will explain in this section. The goal of building a system that helps police officers to predict more accurately if a car contains contraband is a fairly easy problem that we will address in the Modeling phase of this report. The hardest goal to accomplish is to measure the fairness of the decisions made by the police and by our model.

In order to approach this problem, we will explain our whole analysis and propose our model by always having in mind the following two principles:

Principle 1: Acknowledge that statistical evaluation is limited to finding racial,
Ethnic and gender disparities that are indicative of racial and ethnic and gender bias but that, in the absence of a formal procedural investigation, cannot be considered enough evidence.

Principle 2: Always outline the assumptions and limitations of our approach transparently
so that the public and policy makers can use their judgment in drawing conclusions
from the analysis.

Studying Discrimination

The requirement of this project regarding discrimination stated that ”[n]o police sub-department should have a discrepancy bigger than 5% between the search success rate between protected classes (race, ethnicity, gender)”. This is a common metric when studying discrimination. It is normally called *hit-rate* and the assumption is that, for example, if a race has a success rate that is significantly lower from another one, it suggests that there might be some kind of bias in the search criteria for that race in comparison to the other one.
Further considerations about this method and other methods could also be found in the early Connecticut Racial Profiling Reports (CRPR) about these same stop-and-search policies that are avalible online.

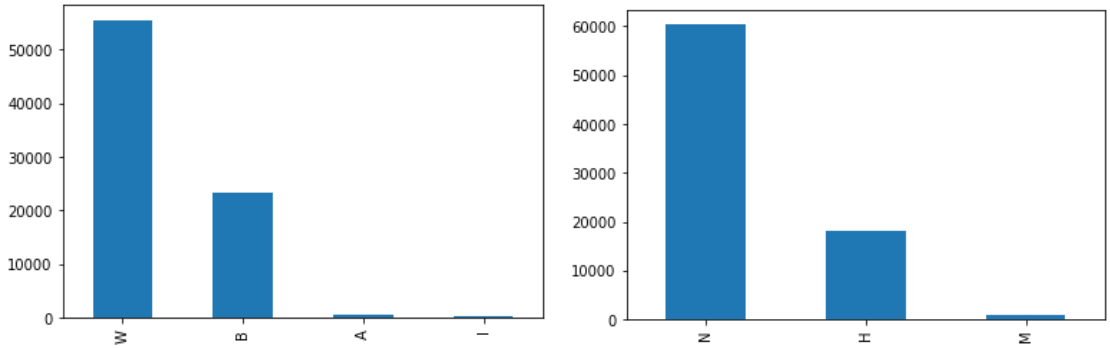
In CRPR of 2016, the main disadvantage of the hit-rate analysis (a post-stop method) is very well explained:

“The disadvantage of post-stop analysis is the small sample size when considering vehicular searches. In many cases, one is unable to estimate the model at the department level because of this issue (...). In addition, as more data is collected there is an increased ability to apply these tests.”

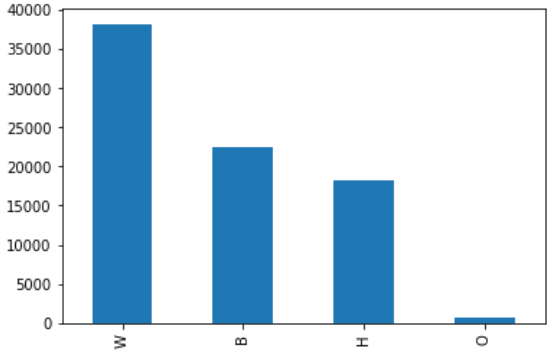
As an illustrative example, consider a department with only 100 searches, 99 of them were white drivers and the other one was a black person. Now if it is found contraband in 30% of the searches for the white people and the black person is found to have contraband, we cannot conclude that white people are discriminated because they do not have the same 100% hit-rate of the black people. In this case, we say that the test is not statistically significant due to sampling size.
In our method, we only considered those comparisons between two classes where both of them had a size of at least 15 observations.

In our case, we have a considerable amount of data available to apply this method in most cases, but we still have some challenges besides this: (1) the classes are not all represented equally in sample size, especially if we analyze each department separately and (2) there is probably a problem of distinction in identifying between race and ethnicity (e.g most people in the dataset have no defined ethnicity as shown in the previous subsection).

Therefore, we decided to merge the ethnicity and race in a single variable with the possible values of: *Hispanic, White*, *Black* and *Other*. With this, we turned two fragmented and unbalanced variables into only one which is more reliable and more balanced. This does not allow for a separate analysis for the discrimination of Asians, Indians and Middle Easteners, but their population size would not allow for a statistically significant analysis anyway. Furthermore, with this new variable we can maybe find some statistically significant results for this minority group all together.



Race(Left): W- White; B- Black; A-Asian/Pacific Islande; I-Indian America/Alaskan Native
Ethnicity (Right): H-Hispanic; M-Middle Eastern; N-Not Applicable



Race_Ethnicity (New Variable): W-White; B-Black; H-Hispanic; O-Other

Testing Discrimination

In order to study discrimination **within each department**, we conducted t-student tests on the hit-rate of the class (gender or race/ethnicity) with higher hit-rate (of that department) and all the lower ones. A t-student test is normally made when we want to see if the difference of two ratios are due to chance. We will spare the technical details in this section and more can be found further in this report. We are doing this test under the previous stated assumption that all the classes, if treated fairly, should have equal hit rates so, if we subtract them, the result should be zero. If the result is not zero, the t-student test gives the probability that the difference of the hit-rates is not due to chance. As asked in the client requirements we will only consider hit-rate differences higher than 5% and we will only consider probabilities of discrimination above 5%-

Referring to the results in Annex A, we can note significant discrepancies in the following departments according to:

- Ethnicity/race (48 problematic):
    - Bethel, Bethel, Bloomfield, Bloomfield, Brookfield, CSP Troop C, CSP Troop C, CSP Troop D, CSP Troop F, Cheshire, Cheshire, East Hartford, Glastonbury, Glastonbury, Greenwich, Greenwich, Groton City, Groton City, Hartford, Milford, Naugatuck, New Britain, New Haven, Newington, North Haven, Old Saybrook, Old Saybrook, Plainville, Rocky Hill, SCSU, Stamford, Stamford, State Police, Torrington, UCONN, Vernon, Wallingford, Waterbury, Waterbury, Waterford, West Hartford, West Hartford, Westport, Wethersfield, Willimantic, Willimantic, Windsor Locks and Yale
- Sex (16 problematic):
    - Bristol, CSP Troop F, Canton, East Hampton, East Windsor, Groton City, Hartford, Madison, Monroe, Plymouth, Putnam, Ridgefield, Thomaston, Windsor, Winsted and Yale

Conclusion

In this section, we will introduced a statistical to measure the discrimination in each department. It is important to repeat that this information alone is not enough to conclude that some Departments are inherently racist or sexist. A statistical test like this is not able to distinguish between discrimination on the basis of race and discrimination on the basis of characteristics that are related to race and that are not available in the datase. Some further in-person investigations should be made and demographic data should be collected before concluding that there are no other variables (considered not to be protected classes) that determine the differences in hit rates found in some departments.
Despite this, we will still suggest a model that tries to eliminate this statistical differences, but that should only be employed if all of the hypothesis about the problematic departments seem to persist (or are even confirmed) after further investigation.

Modeling

Model expected outcomes overview

Model specifications

In order to create a system that finds the “largest possible amount of contraband”, we need to train a machine learning model that receives data from an observed car and that outputs a ***Yes*** if it predicts the existence of contraband inside it and ***No*** otherwise. The data that the model will receive will be of the same format as the one described in the Data Analysis section. Besides the variables already explained, we derived some more features from those in order to improve the model.

Feature Engineering

After multiple tries to create useful features related to time, location and demographics for improving the model, we came to the conclusion that the best model was the one which used only 4 of original (features with some preprocessing):

- SearchAuthorizationCode
- Department Name
- InterventionLocationName
- StatuteReason

All the attempted derived features will be explained further, since they bear no importance in the final mode we are explaining here. It is still important to have them registered in this report, so that the choice of not including them in the model is shown to be done on purpose.

Data Preprocessing

Missing Values

For the InterventionLocationName, we analyzed the target distribution for the missing values and found a 66.7% of true class, which is high, so we created a Location called "*unknown*" that could help predict positive classes if there are missing values in new observations.

For the StatuteReason and SearchAuthorizationCode, there is already a class called "*Other*" so we impute that for missing values for the test set, and we drop the problematic observations for the training set.

String manipulation

For the Department Name, we removed useless spaces and lower cased all of the department names in order to merge inputs with different caps or with extra spaces into the same category.

CatBoost/ Target Encoding

After trying different category encoders as one-hot encoders, ordinal encoders,target encoders and CatBoost encoders the combination that produced the best results was the CatBoost encoder for Department Name and InterventionLocationName and the target encoder for SearchAuthorizationCode and StatuteReason.

Machine learning

The model we chose LightGBM, a (A Highly Efficient Gradient Boosting Decision Tree) proposed by Microsoft. The specifications of this model will be provided in the annex B.The model we chose LightGBM, a (A Highly Efficient Gradient Boosting Decision Tree) proposed by Microsoft. The specifications of this model will be provided in the annex B.

Expected results without bias correction

After we split the original set into a train and test sets according to the date (train: before 2018; test: 2018) we trained our model on the training set and evaluated on the test set.
The output of the models are probabilities that represent the probability that each observation belongs to the true class. In order to output a **Yes (corresponding to 1)** or **No (corresponding to 0),** we have to define a threshold that assigns one of these classes to each probability. Choosing 0.5 would be the naive choice, but we can make a better one in order to meet the requirements of the project.
First we chose to meet the first criterion ("A minimum 50% success rate for searches"). This criterion corresponds to the definition of **precision**, so we iterated through all the k in [0,1] and found that that k had to be, at minimum, 0.17. In order to give some margin of error to account for some more variation in the production data (and to some possible effect in the bias correction phase) we increased the minimum precision to 60% and the resulting k was of 0.34. Then, given that k had to be higher than 0.34, we maximized it in order to find "the largest possible amount of contraband found, given the constraint[...] above". This goal is defined by **recall**, since it measures the ratio of how much contraband we predicted to how much contraband actually existed. The resulting recall of this maximization was 0.78

Correcting bias

In this section we will explain with more technical detail what we already explained in the discrimination sub-section of the **Dataset Analysis** section.

"*In order to study discrimination in each department, we conducted t-student tests on the hit-rate of the class (gender or race/ethnicity) with higher hit-rate and all the other ones. We are doing this test under the previous stated assumption that all the classes, if treated fairly, should have equal hit rates so, if we subtract them, the result should be zero*".

In order to make a t-student test, we have to compute t and then look at the t-student distribution in order to obtain the wanted probability.

$$t = \frac{\text{statistic} - \text{hypothesized value}}{\text{estimated standard error of the statistic}}$$

So this will be our test statistic that we will compute for each department: the difference between hit rates between a class (race/ethnicity or gender) with low hit rates and the class with maximum hit-rates within a department.

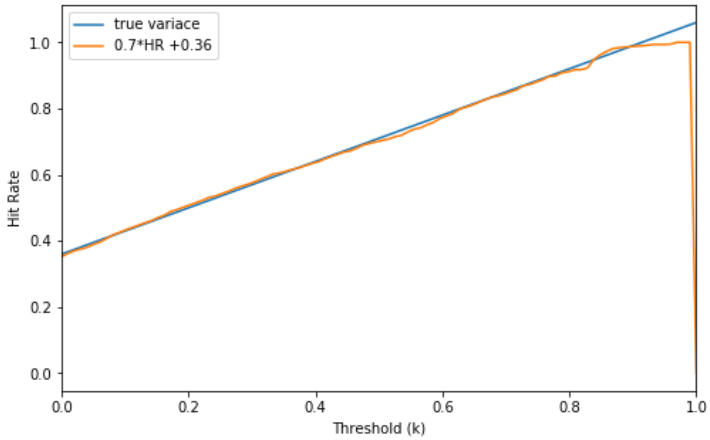
Statistic = M_i - M_{max} , where i is each class within the department that does not have the maximum hit-rate.

For the estimated standard error of the statistic, we use the classical formula that accounts for the different variances within each class and their population size, which is really important for our dataset which is really imbalanced.

$$\sigma_{M1-M2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

And the hypothesized value is that the difference between hit-rates is zero, and so, "*if the result is not zero, the t-student test gives the probability that the difference of the hit-rates is not due to chance, and is statistically significant. As asked in the client requirements we will only consider hit-rate differences higher than 5% and we will only consider probabilities of discrimination above 5%*". For this section, we remade the t-student test just for the training set, and applied the corrections to the test set, in order to simulate the real world scenario that we will be facing.

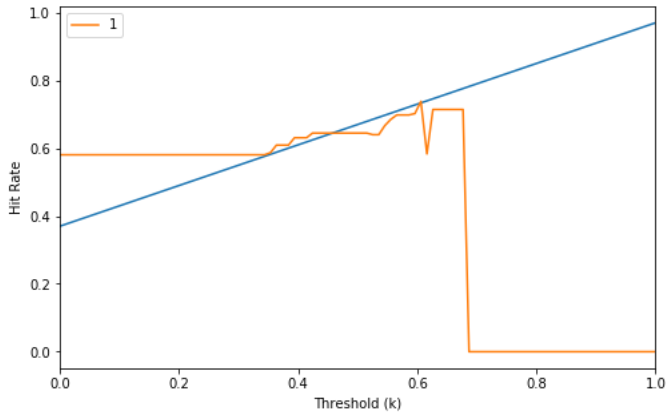
After obtaining these probabilities, we try to correct the prediction of the classes that had a probability higher than 10% of discrimination by adjusting the threshold for those classes.
In order to estimate how much to change the threshold (in order to approximate the discriminated classes to the maximum hit rate within that department), we made a plot of the variation of the Hit Rate with the variation of the threshold for the whole dataset and used those results as an approximation. The results show that the hit rate follows a line that is represented below:



For example, if we have a difference in hit-rates for Male and Female of 0.2 in Ansonia, Female being the lowest with 0.5, and statistical discrimination is found for the Female class, we should increase the threshold for the female class by 0.2/0.7 (0.7 being the slope of the aproximated line).

Results

After finding the departments that had problems with either race/ethnicity or sex and saving the differences of hit-rates found in the results of the model in the training set, we saved these results in 2 files.
After that we made the same analysis in the results of the model for the test set rand then applied the corrections explained previously. We were able to reduce from 26 to 17 departments (from adjusting the ethnicity/race) and from 18 to 14 (from adjusting the sex).
As we see, it will not be possible to meet the second requirement completely, either because of the lack of data for some departments or because of the intrinsic properties (or maybe these properties also vanish with the increasing of the observations) of some departments. For example, there are some departments for which we could increase the threshold more, in order to increase the hit-rate, however, this would mean that no people of that class would ever be ever searched.



An example of a problematic class, for which we have not sufficient evidence to understand this behaviour

Using the whole original set

As we saw in the previous sections, training the part of the model that tries to cancel the differences in hit-rates is hard, mostly because of the lack of data for some departments. We also will receive data points following the timestamps of the test set, so we will train a final model on the whole dataset in order to augment the data the model is trained in, but also for it to generalize better for the data points we will receive. This might increase the uncertainty of the results of the deployment, but we made sure we

Alternatives considered

We tried some feature engineering for many new features, but none of them improved the model. The following list shows the tried ideas:

- Different encoders (binary, ordinal)
- Scaler (minmax, standard)

- Combinations of all these features. Exhaustive search of combinations of all features until N=6
- Location features: coordinates (latitude longitude), drug-related mortality, population and density
- Time related features: day of year, hour, day of week (all of these in sin, cosine and absolute value form),

References

A New Look at Racial Profiling: Evidence from the Boston Police Department
Kate L. Antonovics, Brian G. Knight -

> "The power of this insight is that, even if black and white motorists differ along dimensions other than race, the probability of guilt conditional on search will still be the same for all groups. This is critical since it is generally impossible to distinguish between statistical discrimination on the basis of race and statistical discrimination on the basis of characteristics that are correlated with race but that are unobserved to the econometrician (see for example, Altonji and Pierret (2001) and Dharmapala and Ross (2004)). "

https://www.ccsu.edu/imrp/Publicatons/Files/May%202016%20Connecticut%20Racial%20Profiling%20Report.pdf

Annex A - Department Analysis

Race_Ethnicity Discrimantion probabilities

Department Name	Black	Hispanic	Other	White
Bethel	0	0.107769	0	0.044622
Bristol	0.104196	0	0.476923	0
Brookfield	0.208452	0.172647	0	0
CSP Headquarters	0.164803	0.096232	0	0
CSP Troop A	0.187523	0.165115	0.096067	0
CSP Troop B	0.044163	0.186273	0	0
CSP Troop C	0	0.130634	0.078606	0.176277
CSP Troop E	0.141977	0.199055	0	0.163263
CSP Troop F	0.053833	0.109528	0	0
CSP Troop G	0.12413	0.133454	0	0.021041
Danbury	0	0.074911	0.147436	0.065422
Darien	0	0.111342	0	0.06497
East Hartford	0.080809	0.140528	0	0
Hartford	0.104531	0.101582	0	0
Newtown	0.056572	0.171258	0	0

Norwich	0.03421	0.112452	0	0
Old Saybrook	0.065574	0.287796	0	0
Plainville	0.179128	0	0	0.072049
Plymouth	0.038378	0.107858	0	0
Ridgefield	0.083333	0	0	0.180233
Rocky Hill	0	0.146479	0	0.081661
SCSU	0.335948	0	0	0
Seymour	0	0.119658	0	0.078984
Simsbury	0.25	0	0	0.228723
South Windsor	0	0.152679	0	0.023629
State Police	0.093669	0.088041	0.132576	0
Torrington	0.067515	0.10184	0	0
UCONN	0.020408	0	0.352941	0.176617
Waterbury	0.033371	0.123347	0	0
Watertown	0.197205	0	0	0.087662
West Hartford	0.098016	0.102504	0.360635	0
Willimantic	0.055336	0.106149	0	0
Wilton	0	0.206548	0	0.127944
Windsor Locks	0.102999	0	0	0.14189
Woodbridge	0.162474	0	0	0.068134
Average	0.088244	0.101591	0.049854	0.051518

Sex Discrimantion probabilities

Department Name	F	M
Bristol	0.119329	0
CSP Troop F	0	0.101009
Canton	0.140827	0
East Hampton	0.14016	0
East Windsor	0.204261	0
Groton City	0	0.135618
Hartford	0	0.102086
Madison	0	0.310102
Monroe	0.105988	0
Plymouth	0.17705	0
Putnam	0	0.273684
Ridgefield	0.102904	0
Thomaston	0.236559	0
Windsor	0.145524	0
Winsted	0.452055	0
Yale	0	0.17412
Average	0.114041	0.068539

Annex B

The following specifications are the hyperparameters used in the LGBM model:

```
LGBMClassifier(boosting_type='gbdt', class_weight=None,
               colsample_bytree=1.0, importance_type='gain',
               learning_rate=0.1, max_depth=-1,
               min_child_samples=20, min_child_weight=0.001,
               min_split_gain=0.0, n_estimators=1000,
               n_jobs=-1, num_leaves=31, objective=None,
               random_state=None, reg_alpha=0.0,
               reg_lambda=0.0, silent=True, subsample=1.0,
               subsample_for_bin=200000, subsample_freq=0)
```