Client Requirements

Summary

Every time a car is stopped by Police, Police Officers have to decide if they should search the vehicle for contraband or not. Their policy for deciding if a stopped vehicle should be searched or not has received a lot of complains. People complain about Police Officers tendency to decide to search vehicles based on people's background[1] (age, gender, race...).

Awkward Problem Solutions™ developed a model to help deciding when a vehicle that was stopped by Police Officers should be searched or not for contraband, without harming people because of their background.

This report describes the analysis made on historical data of Police Departments that have information about past stopped vehicles. This data has information about if the vehicle was searched or not, if yes if contraband was found or not and other informations about the intervention. Data analysis has the objective to understand indicators of Contraband and if Police Officers are searching vehicles taking in account peoples' background. The report also describes the model developed to predict whenever a stopped vehicle should be searched for contraband or not.

Requirements Clarifications

An analysis to the historical data about stopped and searched vehicles should be done in order to find out what seems to indicate contraband and if Police Departments have any bias against People of certain backgrounds.

Also, a fair model to decide when a vehicle need to be searched for contraband or not should be developed. There are some requirements that the model from *Awkward Problem Solutions*™ should achieve. The model will receive information about stopped cars and return back to the Police Officer the information about if he should search the vehicle for contraband or not. Like mentioned before, the model should not have bias against people of certain backgrounds. To describe the requirements in a quantitative way, let's define success rate for searches:

$$precision = 100 * \frac{true\ positives}{true\ positives\ + false\ positives}\ \%\ \ (1)$$

The model should have at least 50% of precision[1]. Also, no police department should have a difference in precision bigger than 5 between each protected group[1]. This means that each police department shouldn't have an amplitude[1] bigger than 5 between precisions of protected groups.
Taking this requirements in account, the number of true positives[1] should be maximized.

[1] Annexes-Business questions technical support

Dataset Analysis

General analysis

The Dataset, with 15 variables, has historical data for 5 years with data about 2268887 unique interventions. For each stopped vehicle it has the following information:

Category	Column Names	Description
Target[2]: Police Officer Prediction	VehicleSearchedIndicator	If the vehicle was searched for contraband or not.
Target: Real Value	ContrabandIndicator	If the vehicle was searched for contraband, if contraband evidences were found.
Time/Date and Location	InterventionDateTime InterventionLocationName	Date/time and location of the intervention, respectively.
Police Information	Department Name	Police Department Name that is doing the intervention.
	ReportingOfficerIdentificationID	Police Officer who is in charge of the intervention.
Intervention Information	StatuteReason InterventionReasonCode	Reason to stop the vehicle and its code, respectively.
	search_authorization_code	Authorization given to the Police Man to search the vehicle, if applicable.
Vehicle Driver Personal Information	subject_sex_code SubjectRaceCode SubjectAge SubjectEthnicityCode	Vehicle Driver personal information (gender, race, age, and ethnicity, respectively).
Vehicle Driver Information	TownResidentIndicator Resident Indicator	If Vehicle Driver is a town/state resident or not.

Table 1: Dataset Variables Description.

Variables on Table 1 are divided by logical categories so its analysis can be easier.

[2] Annexes-Business questions technical support

Target Category

Target Category represents variables that act as target (remaining categories act as features[3]) in a machine learning perspective. VehicleSearchedIndicator is the Police Officer prediction and ContrabandIndicator the real value for contraband evidences in the vehicle being present or not. Note that we can't spot false negatives[3] or true negatives[3] on the dataset because if the vehicle was not searched, we will never know with 100% sure if the vehicle has contraband evidences or not. But the rule "everyone is innocent until the opposite is proved" will be followed and every vehicle that was not searched will be considered a true negative[3].

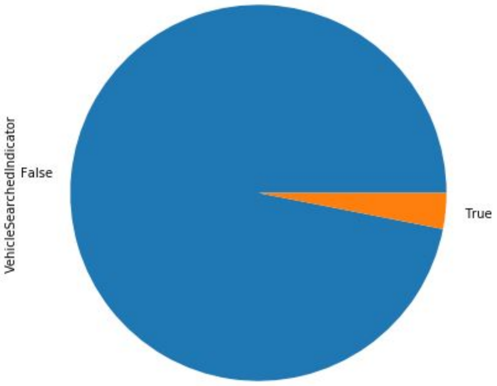


Figure 1: VehicleSearchedIndicator values distribution, True if the Vehicle was searched and False if not.

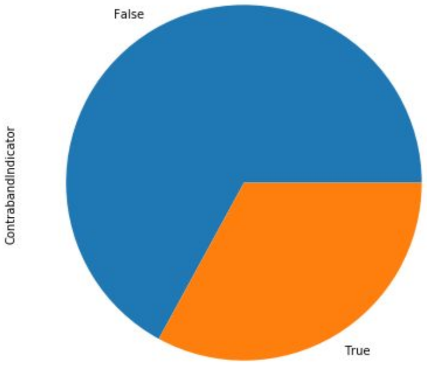


Figure 2: ContrabandIndicator values distribution, when VehicleSearchedIndicator value is True. True if contraband evidences were found and False if not.

From pie charts on figures 1 and 2 we can conclude that the percentage of evidence found on total vehicles stopped is very low. Also, for the vehicles searched, the percentage of evidence for contraband found is lower than half. There is also a small percentage of vehicles that were not searched but had contraband evidences that is not represented in the figures above. Those cases were so obvious that the police officers didn't search the vehicles.

Time/Date and Location Category

We have historical data between 2013-10-01 and 2018-05-16. Years with more data are 2014, 2016 and 2017 (descending order). The distribution of the number of stopped vehicles is uniform for weekdays and months. The distribution of hours is in Figure 3.

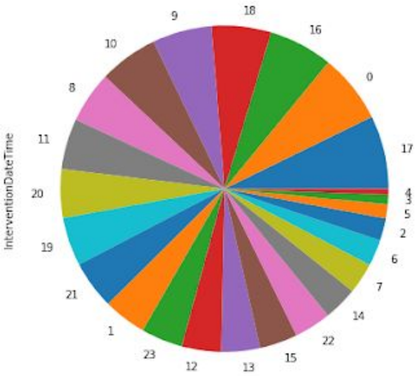


Figure 3: Distribution of hours.

[3] Annexes-Business questions technical support

For location of the interventions, after cleaning the data we have 827 different locations but just 218 of them have more than 100 interventions registered in 5 years. Top 3 locations with more interventions are Newhaven, Stamford and Hartford (descendent order).

Police Information Category

This category has information about the Police Department name that is doing the intervention (Department Name). After cleaning the referred variable we came to conclusion that we have 120 distinct department names, almost every one with more than 100 interventions. Top 3 with more interventions are State Police, New Haven and CSP Troop C (descendent order).

This category also has the Id of the Police Officer (ReportingOfficerIdentificationID) that is doing the intervention. We have 8593 different registers of Police officers. The one with more records has 8183 distinct interventions.

Intervention Information Category

This Category has the information collected about the intervention. It has reasons to stop the vehicle (StatuteReason). Top 3 reasons to stop the vehicle are Speed Related, Defective Lights and Registration (descendent order). We also have the code given to stop the vehicle (InterventionReasonCode) represented on Figure 4.

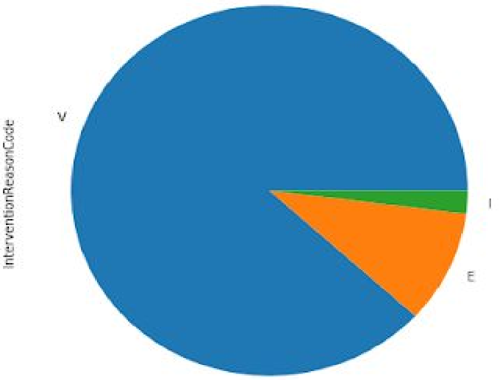
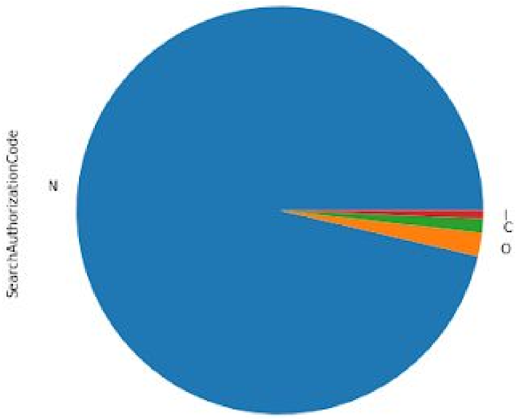


Figure 4: Values distribution of InterventionReasonCode.
V: Violation; I: Investigation ; E: Equipment;

On this category we also have the Authority of Officer to search the vehicle (search_authorization_code). This variable has the prior perception, before the vehicle be searched for contraband, if the car is suspicious or not. Because of this this, this variable can have an implicit bias for people of certain backgrounds. This variable is divided in 4 values and it has the following distribution:



N-Not Applicable
C-Consent
I-Inventory
O-Other:Probable Cause, Reasonable Suspicion, Plain View Contraband, Incident to Arrest, Drug Dog Alert, Exigent Circumstances

Figure 5: Distribution of search_authorization_code values.

Vehicle Driver Personal Information

Variables on this category are all related with personal information about the driver. All this features should not be used to train a machine learning model because a decision should be fair and not based on personal information. On the images bellow it is possible to see the values distribution of each variable.

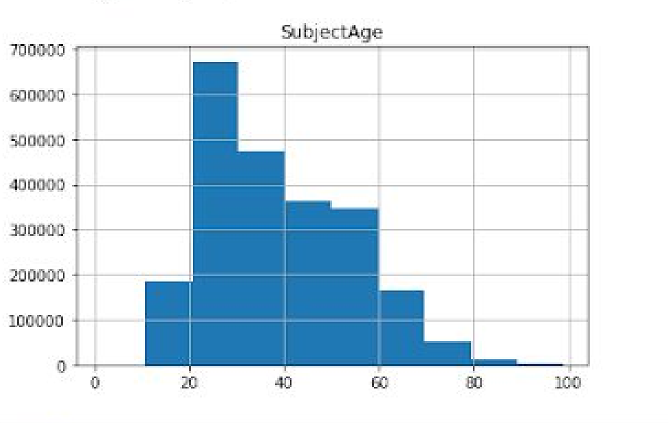


Figure 6: SubjectAge distribution.

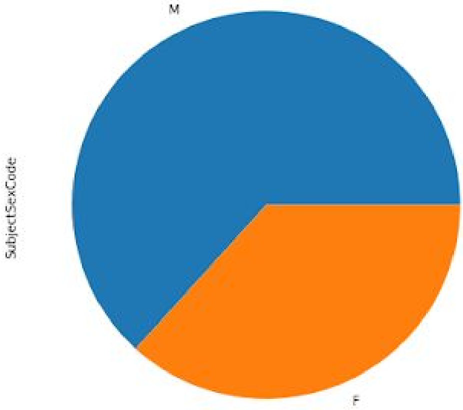
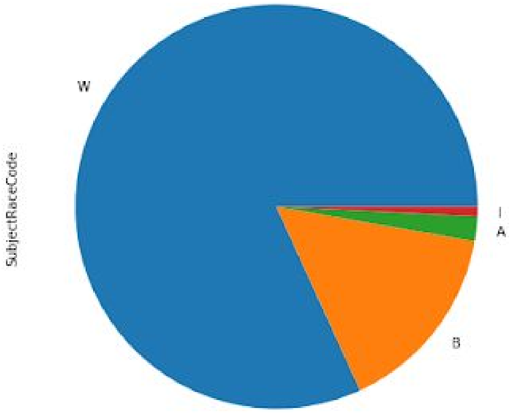
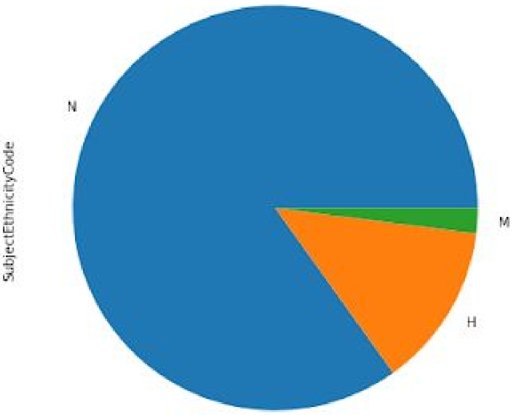


Figure 7: SubjectSexCode distribution. M: Male; F:Female

W-White
B-Black
I-Indian America/Alaskan Native
A-Asian/Pacific Islander
U-Unknown

Figure 8 (right): SubjectRaceCode values distribution;





H-Hispanic: H
M-Middle Eastern: M
N-Not Applicable

Figure 9 (left): SubjectEthnicityCode values distribution;

From figures 7 to 9 we can conclude that we have protected groups that are very low represented in our database. Also, for ages distribution, from figure 6 it looks like the range is between 10 and 100, the minimum age registered id 1 and me maximum is 99. This is a bit weird because people under 16 are not allowed to drive and we are not expecting to see a person with more than 90 driving. Data out of range [16, 90] probably is incorrect data.

Vehicle Driver Information

This information is about the driver but is not strictly personal information. Is the information about if the driver is a resident of the town or state where the intervention is being done. If they are residents maybe the reason for them to be in that location is just because they live nearby and no other suspicious reason. Or, in other hand people can do this kind of crimes in their towns so people will not suspect. For this reason, this maybe can be a good indicator to spot contraband without harming the driver because is not a personal information like zip-code. On figures below we can see the distribution of the variables for town indicator (Figure 10) and state indicator (Figure 11). True for when driver is a resident and False for when he is not.

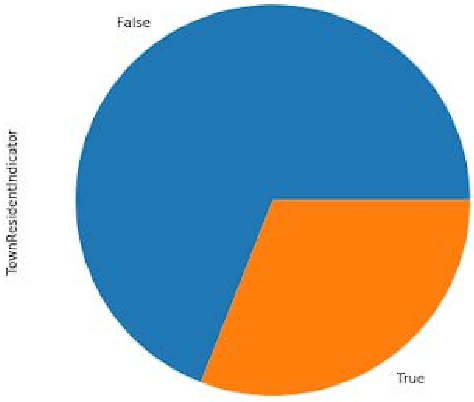


Figure 10: TownResisentIndicator distribution.

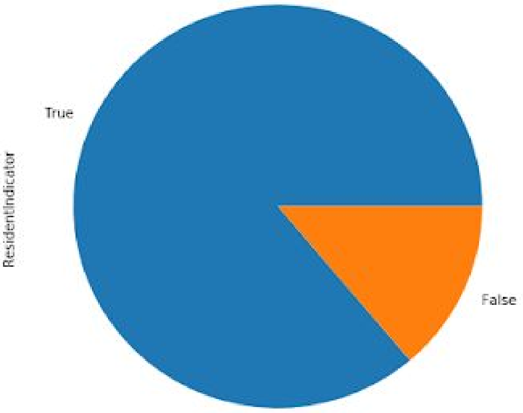


Figure 11: ResisentIndicator distribution.

Business questions analysis

Let’s start to analyse what seems to indicate contraband.

After some data cleaning and encoding we can calculate the correlation[4] between target (the real value), ContrabandIndicator, and variables from remaining categories (features), represented on table 1 and results are represented on Figure 12.

[4] Annexes-Business questions technical support

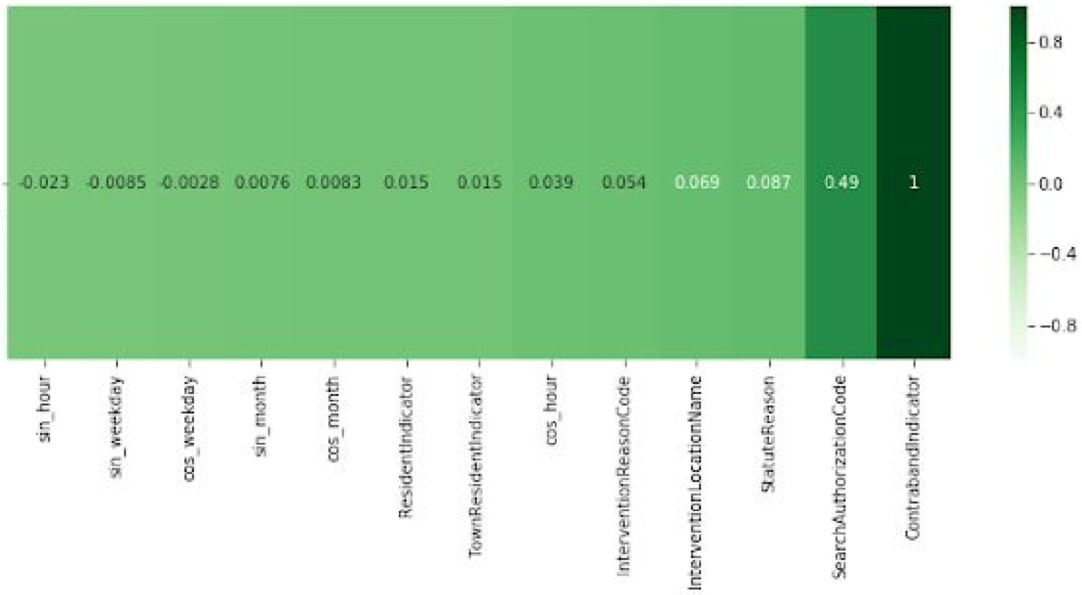
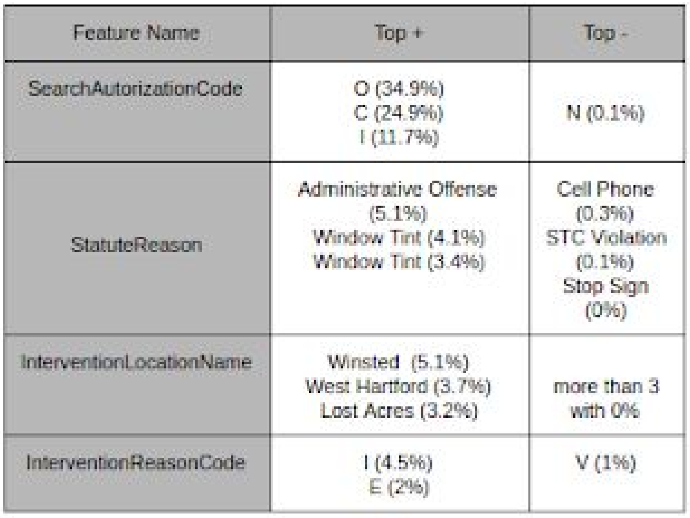


Figure 12: Correlation with ContrabandIndicator

From Figure 12 we can conclude that the feature that is more correlated with target is SearchAutorizationCode. StatuteReason, InterventionLocationName, InterventionReasonCode and hour are also variables that have a considerable correlation with the target. Let's take a look to the values for these variables that have a highest and the lowest percentage of contraband found per number of vehicles stopped:



Feature Name	Top +	Top -
SearchAutorizationCode	O (34.9%) C (24.9%) I (11.7%)	N (0.1%)
StatuteReason	Administrative Offense (5.1%) Window Tint (4.1%) Window Tint (3.4%)	Cell Phone (0.3%) STC Violation (0.1%) Stop Sign (0%)
InterventionLocationName	Winsted (5.1%) West Hartford (3.7%) Lost Acres (3.2%)	more than 3 with 0%
InterventionReasonCode	I (4.5%) E (2%)	V (1%)

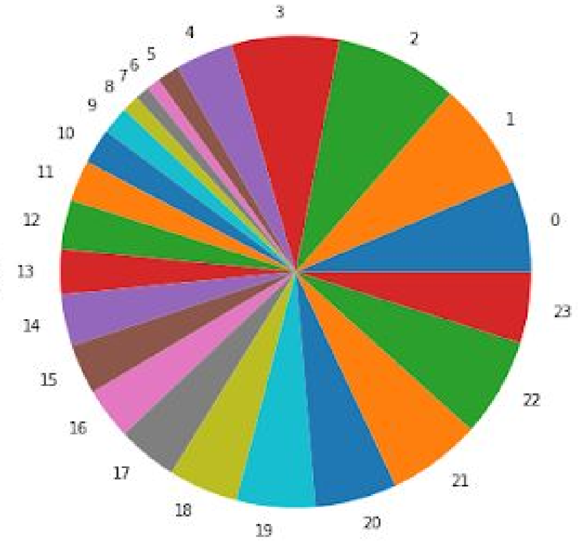


Table 2: Values with highest and the lowest percentage of contraband found per number of vehicles stopped.

Figure 13: Percentage of contraband found per number of vehicles stopped for each hour.

To identify if Police Departments are biased against people of certain backgrounds, every department should be analysed in an individual way.

For every Police Department, differences between precision (1) were calculated for every possible combination of protected groups[5] that had more than a threshold of records. This threshold, 100, was defined because we need to have a considerable number of records to evaluate if the police department is fair or not.

All the differences calculated for every police department, between every protected group (protected groups with more than 100 records), are represented on figure 14 bellow as an histogram.

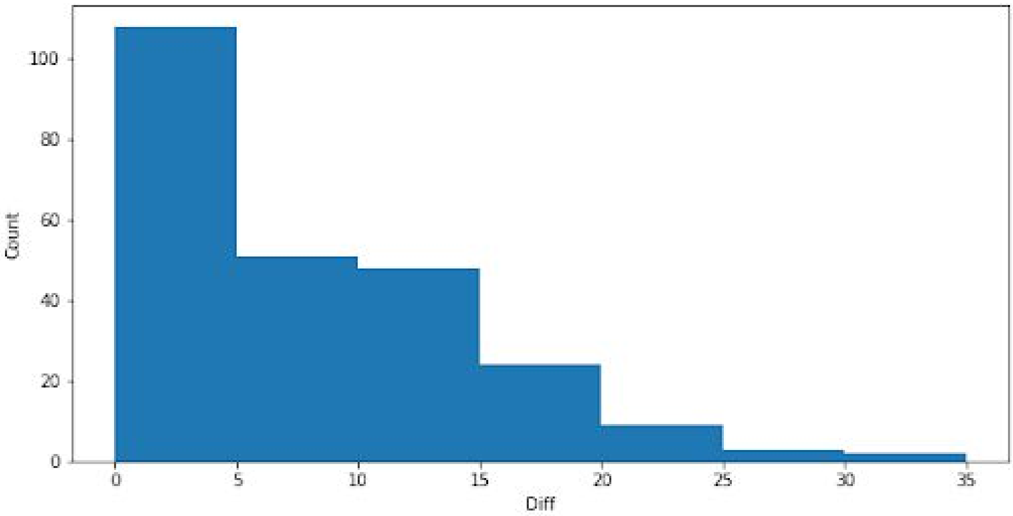


Figure 14: Frequency of differences in precision between protected groups from the same department name. All the differences for protected groups (registered by the same department)) with more than 100 records, for every department, are represented.

From Figure 14 we can conclude that we have a lot of protected groups with a difference bigger than 5 for precision between them. As a remender, this difference is calculated between records from the same department, there is no mixing of records from different departments.

For each protected group, average of the precisions calculated for each department name were calculated. The protected group with highest precision on average, 45.5%, is white women from a young age with no ethnicity registered by the police officers. The protected group with lowest precision on average is Hispanic white women from a young age with 4.3%.

[5] Annexes-Business questions technical support

Amplitude (maximum precision - minimum precision) between precisions of protected groups for each police department also was calculated. Police Department name with highest amplitude was Wallingford (34.4 of amplitude) and the lowest Middletown (0.5 of amplitude). On average, amplitude was 12 and there are 30 police departments with an amplitude bigger than 5.

Same procedure applied to police department names can be also applied to SearchAuthorizationCode in order to find out if this feature is a source of bias as expected.

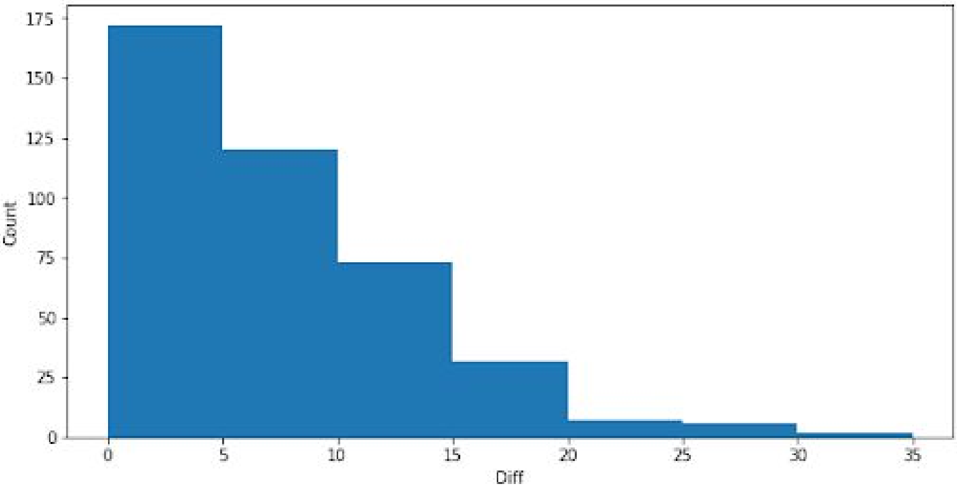


Figure 15: Frequency of differences in precision between protected groups from the same authorization code. All differences for protected groups (registered for the same authorization code) with more than 100 records, for every code, are represented.

From figure 15, we can conclude that SearchAuthorizationCode has a bias in it. The reason for that is because this feature represents the perception (if the intervention is suspicious or not in terms of contraband) of the Police Officer before searching the vehicle. So, as we conclude before, police department has a bias against protected groups. If Police Officers have a bias against people of certain backgrounds, their prior perception of the intervention, before searching the vehicle will also be biased.

Conclusions and Recommendations

From this analysis to the historical data of interventions we can conclude that variables like SearchAuthorizationCode, StatuteReason, InterventionLocationName, InterventionReasonCode and hour are good indices if the vehicle should be searched for contraband or not.

Also, as required, an investigation was done to police departments to understand if they have a bias against some protected groups. Findings are that some police departments are biased. Police departments that have a amplitude bigger than 5 (for protected groups that have more than 100 records registered by department name) in precision for protected groups are: Wallingford, State Police, Vernon, Glastonbury, Waterbury, wilton, CSP Troop G, Norwich, CSP Troop C, CSP Troop E, CSP Troop A, West Hartford, Enfield, Danbury, New Britain, West

Haven, Berlin, Willimantic, Westport, East Hartford, Trumbull, CSP Troop H, Norwalk, Plainville, wethersfield, Stamford, Bridgeport, Manchester, Stratford and New Haven.

But, we have a lot of protected groups that have less than 100 records for each department name. We have 4686 protected groups by department name with less than 100 records and just 1772 with 100 or more records. This means that we are not spotting bias for a lot of protected groups by department name that are low represented in our dataset.

We chose this threshold because it is the number of records that we consider enough to evaluate if a police department is biased or not. Because we need to have a considerable number of records for a protected group registered by police department to be sure if they are being fair or not for the correspondent group.

The same analysis was done for police officers, but with a threshold of 25 records and we came to conclusion that the following police officers with the correspondent Id need further investigation: 1000003196, 1000002101, 199, 10, 100000230, 596, 1000002072, 790642042, 1000002164, JMK0326, 30233, 1000002585, 625, KKC0268, SEM0240, 179, 256, 49, 1016, 6507, 1000002747, 1000002608 and 1083.

Modeling

Model expected outcomes overview

When a Police Officer stops a vehicle, police officer should send to our model all the information collected by him. Our model will respond back with a prediction if the vehicle should be searched for contraband or not. As explained before, this model should not be biased against certain protected groups.

As we can see on table 1, we don't have a lot of information (features) to train the model. Variables from Police Information and Vehicle Driver Information categories will not be used to train the model because it will induce a lot of bias. With this, we don't have a lot of remaining variables and the ones we have are not much correlated with target as we can see on figure 12. To add to this, the variable that is must correlated with target, SearchAuthorizationCode, has some implicit bias on it.

With the exposed information we are not expecting to achieve all the requirements for the model because in order to reduce bias we will be reducing the final score of the model and vice-versa.

Model specifications

The type of supervised model that we want to train is a binary classification because our target (variable that we want to predict) have only two categorical values, True and False, that can be encoded as 0 or 1.

Before anything else, we need to start by splitting the dataset in training and test set. This should be done in a stratified way, this means that training and test set should have the same percentage of contraband found per total number of records. One of our features is a timestamp but our model in not a time series model, it just looks to variables like weekday, month or hour so we can shuffle the dataset and split it randomly.

After splitting the dataset, we divide our new datasets in features X, and target y. Target is the column ContrabandIndicator and features are the remaining columns with exception of VehicleSearchedIndicator.

We will make use of pipeline objects so the conversion of a fitted pipeline to a file and its reproducibility can be easier.

In the fit process, this pipeline will recebe X and y from the training dataset. Our model doesn't use features from category personal information or police information (Table 1). Our pipeline will drop these features.

The first steps of this pipeline is data cleaning. Variables that need to be cleaned before anything else are InterventionLocationName, StatuteReason and InterventionReasonCode.

Data cleaning process for these variables are almost the same. Those variables have categorical values and their values are strings. We start by converting the values to lowercase and removing spaces. For StatuteReason and InterventionReasonCode we also remove any non word character and after that, all values with more than 100 records in the training set should be saved in a list in the fit process to be used in the transforming process. Remaining values are replaced by a new value called "others". With this new value called "others" we can deal with new values that can appear in the training set. For InterventionLocationName cleaning is a bit more complex because we have locations that don't have a standardized way to be written. Values with "@" or "/" are splitted in two strings by the symbol referred and only is considered the first string. Also, there are a lot of values ending with "ve", "venue", "st", "street", "dr", "drive", "ville". If a value end with these characters, they are removed from the end of the string. Values like "none", "na", "" and "nan" are replaced by None and values like "othertown" and "yourcity" replaced by others. Also, when "ve", "venue", "st", "street", "dr", "drive", "ville" appear alone they are replaced by "others". After this cleaning process, the same procedure for replacing values with less than 100 occurrences is applied to this feature.

After this is time to clean InterventionDateTime. This feature is a string that has in it information to create a timestamp. It has the day, month, year, hour, minute and second and it is in a.m./p.m format. These stings need to be parsed in order to create a timestamp. With timestamp created it is possible to extract time features like hour, weekday and month. In this case, only hour will be extracted because month and weekday are not very correlated with target as we can see on figure 12. After extracting hour, it need to be converted on its cyclical features.

Now that all features that we will use in our model are cleaned, it is time to move to the encoding part.

We created our own way to encode categorical features for this case. This econding consists on replacing each value by the percentage of contraband found per number of stopped vehicles when this value appears in the training dataset. This encoding is used for InterventionLocationName, StatuteReason and InterventionReasonCode.

ResidentIndicator, TownResidentIndicator, SearchAuthorizationCode and our Target VehicleSearchedIndicator are encoded as boolean features.

ResidentIndicator, TownResidentIndicato and VehicleSearchedIndicator have as values True and False. True is encoded as 1 and False as 0.

SearchAuthorizationCode has as values “N”, “C”, “I” and “O”. “C”, “I” and “O” are encoded as 1 and “N” as 0. This feature is encoded in this way because we want just to tell to the model whenever the police officer has authorization to search the vehicle or not. This is a way of trying to reduce a little bit the bias of this feature.
For the boolean features, when a new value appears in production it will be encoded as 0.

For the referred variables, nulls values are treated as any other value. For boolean features they are encoded as 0 and for the remaining categorical values that are encoded with the percentage of positives for none value. If value none has more than 100 occurrences stays as none and encoded with the corresponded percentage, if it has less, is converted to “others” and encoded with the corresponded percentage.

At this time, we have all features encoded. We will use tree based models. Because of that we can skip the part of scaling features.

Because our percentage of contraband encoding depends on our target, techniques to turn our dataset more balanced are just applied after cleaning and encoding. Before fitting the model, smote algorithm is applied to dataset in order to balance it.

With this done, we can train the model. The model that we choose was AdaBoostClassifier.
We used 50 estimators, a learning rate of 1 and SAMME.R algorithm as hyperparameters for AdaBoostClassifers model.

Analysis of expected outcomes based on training set

Train dataset, as referred before, was splitted in train and test set. Our model was fitted in our training set and tested with test set. From results obtained in our test set we are expecting the following results in production:

-	precision	recall	f1-score	support
0 (negative)	1.00	0.97	0.99	489061
1 (positive)	0.28	0.89	0.43	5668
accuracy			0.97	494729
macro avg	0.64	0.93	0.71	494729
weighted avg	0.99	0.97	0.98	494729
-	true negative	false positive	false negative	true positive
-	476125	12936	630	5038

Table 3: Evaluation of model performance. Remember that 0 is the encoding for value False and 1 for True.

On table 3 it is possible to see the expected results. Remember that 0 is the encoding for value False on the target (ContrabandIndicator) and 1 for True value. F1 macro average of 0.71 was the maximum score obtained taking in account all requirements. According our model, approximately for 3 vehicles searched, one will have contraband evidences.

Let’s now evaluate results for bias. Analysis made on predicted results were the same applied on figure 14.

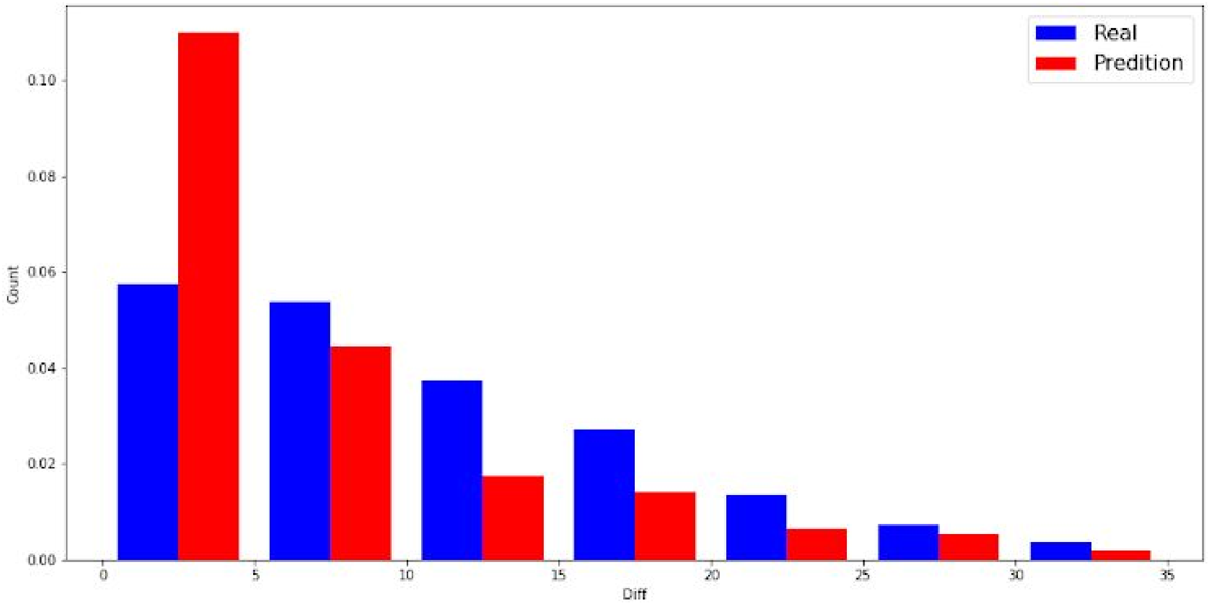


Figure 16: Frequency of differences in precision between protected groups from the same police department. All differences for protected groups (registered for the same police department) with more than 20 records, for every police department, are represented. Results are in percentage and is represented both model prediction (prediction) and police officers prediction (real).

On figure 16 we can compare results for bias between police officers predictions (Real-blue) and model prediction (Prediction-red). Results are in percentage so we can compare real with prediction. Methodology to create this graphic was the same as on figure 14. From this visualization, figure 16, we can conclude that our model reduce differences in precision between protected groups because red bar if bigger than blue bar in range [0-5] and always smaller than blue bar in other bigger ranges. But the requirement of not having a difference bigger than 5 between protected groups by department name in precision is not completely achieved.

Alternatives considered

In order to compare alternatives, we used f1 macro average score obtained from cross validation, applied to each different pipeline.

To encode features InterventionLocationName, StatuteReason and InterventionReasonCode label encoder and one hot encoding was also considered. One hot encoder was discard because it can overfit the model and we are using tree based models so there is no necessity of using it. The doubt was between label encoder and our percentage of positives encoding but with our encoding we had better results.

To encode SearchAuthorizationCode we also tried to apply the percentage of positives encoding but we had better results with the boolean encoding. We also tried to implement a way of penalize police departments with bias against people of protected backgrounds in this feature. The idea was to multiply value 1 already encoded by 1 - amplitude[6]. Like this, value 1 for police departments with bigger amplitudes are not so significant to the model. But the improvements one the results with this penalization were not significant and we decided to not add this extra complexity to the model. We also tried to multiply the value encoded by the precision for each protected group but the results were also not better. To evaluate if applying regularization improved the model we tried to tune these regularization techniques with different thresholds for protected groups and to power the values with different numbers.

Oversampling and undersampling techniques for balance the dataset were also considered but we got better results with smote.

Finally, other tree based models was also considered. Models were Gradient Boosting and Random Forest. But again, we have better results with AdaBoost.

Like referred before, SearchAuthorizationCode is a big source of bias. We trained the model without this feature in order to evaluate the results. F1 macro average score was 0.44. The bad impact in the score was huge so we decide to keep the feature. But if we had removed this variable we would have better results on reducing bias for protected groups.

[6] Annexes-Business questions technical support

Known issues and risks

As referred, not all requirements were achieved. We were not able of reducing completely police departments bias against protected groups, so there will be some bias in predictions. It was not totally removed, it was reduced.

Our model has a better score to predict no contraband evidences in the car than cars who have it. Because of that the model will be better to predict cars who don't have it than ones who have.

Also, our model will have more tendency to predict contraband when the vehicle actually doesn't have any evidences of it rather the opposite, predicting no contraband evidences when the car actually has.

We can also receive new (unseen) values for the variables that were not in the training set that can be good to re-train the model in future, but until there, they will be all encoded as "others" or 0, depending on the encoding.

Model Deployment

Deployment specifications

In order to deploy our model and make it available to Police officers to use it, a trained pipeline should be saved in a pickle file. This file will be loaded in our Api. This Api will serve as a channel of communication between *Awkward Problem Solutions*™ side and police officers side, were police officers can send questions and have their responses back.

Our rest api has 3 endpoints: /predict, /update and /list-db-contents. It is connected to a database were interactions with police officers can be recorded.

Endpoint /predict is of type POST. Police officers should use this endpoint when they want to know if a vehicle they just stop should be searched for contraband or not. In the request, they should send the following information: Unique ID of the intervention, 'Department Name', 'InterventionDateTime', 'InterventionLocationName', 'InterventionReasonCode', 'ReportingOfficerIdentificationID', 'ResidentIndicator', 'SearchAuthorizationCode', 'StatuteReason', 'SubjectAge', 'SubjectEthnicityCode', 'SubjectRaceCode', 'SubjectSexCode', 'TownResidentIndicator'. They will receive back a boolean, if the response is True police officers should search the vehicle, if False they shouldn't. Data sent in the request and the correspondent response is saved in the database. If in the request is send an id that is already registered in the database, the information on database is updated.

Endpoint /update is also of Post type and is for Police officers to send back the true outcome of the search, if the vehicle had contraband evidences or not. This True outcome is registered in the database. If the id of the intervention sent in the request is not in the database an error should be send back to the police officer.

Last endpoint is /list-db-contents and is of type Get. This endpoint returns back a list with information about all interventions, predictions and true outcomes done until the moment since the api is live.

Api was deployed with heroku and the url is the folowing: http://inespessoaldssa.herokuapp.com

Known issues and risks

Unit tests were not done to the Api. New data coming can generate errors that we are not expecting and that our pipeline didn’t covered. Also, common errors like bad requests, no content and internal server error were not converted so when our Api returns a error the reason will be harder to spot for the end user.

Because of this lack of testes and exceptions covered we have the risk of having complains from police officers that our Api is crashing and they don’t know the reason why.

The server where we deployed our Api has a limited size dedicated to our database. If we have a lot of requests coming in is probable that in some point our api stops working because database is full and it is not possible to save more records.

Annexes

Dataset technical analysis

Dataset Overview

Number of records: 2473643
Number of columns: 15
Columns names: 'VehicleSearchedIndicator' 'ContrabandIndicator' 'Department Name' 'InterventionDateTime' 'InterventionLocationName' 'InterventionReasonCode' 'ReportingOfficerIdentificationID' 'ResidentIndicator' 'SearchAuthorizationCode' 'StatuteReason' 'SubjectAge' 'SubjectEthnicityCode' 'SubjectRaceCode' 'SubjectSexCode' 'TownResidentIndicator'
Column types:

VehicleSearchedIndicator	bool
ContrabandIndicator	bool
Department Name	object
InterventionDateTime	object
InterventionLocationName	object
InterventionReasonCode	object
ReportingOfficerIdentificationID	object
ResidentIndicator	bool
SearchAuthorizationCode	object
StatuteReason	object

```
SubjectAge                  float64
SubjectEthnicityCode            object
SubjectRaceCode                 object
SubjectSexCode                  object
TownResidentIndicator               bool
```

Number of nulls that each column has:

```
VehicleSearchedIndicator: 0
ContrabandIndicator: 0
Department Name: 0
InterventionDateTime: 0
InterventionLocationName: 36
InterventionReasonCode: 2
ReportingOfficerIdentificationID: 2
ResidentIndicator: 0
SearchAuthorizationCode: 10
StatuteReason: 507
SubjectAge: 0
SubjectEthnicityCode: 0
SubjectRaceCode: 0
SubjectSexCode: 0
TownResidentIndicator: 0
```

Number of duplicated records: 204756

Number of unique records: 2268887

Columns Analysis

Number of occurrences for each unique value of the boolean column VehicleSearchedIndicator:

```
False    2199008
True       69879
```

Number of occurrences for each unique value of the boolean variable ContrabandIndicator:

```
False    2243241
True       25646
```

Number of occurrences for each unique value of the categorical variable InterventionReasonCode:

```
V     2000365
E      223501
I      45016
no          3
```

Number of occurrences for each unique value of the boolean variable ResidentIndicator:

```
True     1955377
False     313510
```

Number of occurrences for each unique value of the boolean variable TownResidentIndicator:

```
False    1567003
True      701884
```

Number of occurrences for each unique value of the categorical variable SearchAuthorizationCode:

```
N    2186969
O      42123
C      25307
I     14478
```

Number of occurrences for each unique value of the categorical variable SubjectSexCode:

```
M    1435490
F     833397
```

Number of occurrences for each unique value of the categorical variable SubjectRaceCode:

```
W    1856712
B     350133
A      44419
I     17623
```

Number of occurrences for each unique value of the categorical variable SubjectEthnicityCode:

```
N    1925881
H     297619
M      45387
```

Distribution of values for the continuous variable SubjectAge:

```
count    2.268887e+06
mean     3.864314e+01
std      1.495402e+01
min      1.000000e+00
25%      2.600000e+01
50%      3.600000e+01
75%      5.000000e+01
max      9.900000e+01
```

Unique values of the categorical variable Department Name: 'Ansonia' 'Avon' 'Berlin' 'Bethel' 'Bloomfield' 'Branford' 'Bridgeport' 'Bristol' 'Brookfield' 'CAPITOL POLICE' 'CCSU' 'CSP Headquarters' 'CSP Troop A' 'CSP Troop B' 'CSP Troop C' 'CSP Troop D' 'CSP Troop E' 'CSP Troop F' 'CSP Troop G' 'CSP Troop H' 'CSP Troop I' 'CSP Troop K' 'CSP Troop L' 'Canton' 'Cheshire' 'Clinton' 'Coventry' 'Cromwell' 'DMV' 'Danbury' 'Darien' 'Derby' 'ECSU' 'East Hampton' 'East Hartford' 'East Haven' 'East Lyme' 'East Windsor' 'Easton' 'Enfield' 'Fairfield' 'Farmington' 'Glastonbury' 'Granby' 'Greenwich' 'Groton City' 'Groton Long Point' 'Groton Town' 'Guilford' 'Hamden' 'Hartford' 'Ledyard' 'MET DIST WATER AUTHORITY' 'MTA' 'MTA Stamford' 'Madison' 'Manchester' 'Mashantucket Pequot' 'Mashantucket Pequot Police' 'Meriden' 'Middlebury' 'Middletown' 'Milford' 'Mohegan Tribal' 'Mohegan Tribal Police' 'Monroe' 'Naugatuck' 'New Britain' 'New Canaan' 'New Haven' 'New London' 'New Milford' 'Newington' 'Newtown' 'North Branford' 'North Haven' 'Norwalk' 'Norwich' 'Old Saybrook'
'Orange' 'Plainfield' 'Plainville' 'Plymouth' 'Portland' 'Putnam' 'Redding' 'Ridgefield' 'Rocky Hill' 'SCSU' 'Seymour' 'Shelton' 'Simsbury' 'South Windsor' 'Southington' 'Stamford' 'State Police' 'Stonington' 'Stratford' 'Suffield' 'Thomaston' 'Torrington' 'Trumbull' 'UCONN' 'Vernon' 'WCSU' 'Wallingford' 'Waterbury' 'Waterford' 'Watertown' 'West Hartford' 'West Haven' 'Weston' 'Westport' 'Wethersfield'

Top 5 values with more records for the categorical variable Department Name:

```
State Police    320343
New Haven        64399
```

```
CSP Troop C      53626
CSP Troop F      48550
CSP Troop A      45983
```

Number of unique values for categorical variable InterventionLocationName before lowercase and spaces removing: 2504

Number of unique values for categorical variable InterventionLocationName after lowercase and spaces removing: 1556

Unique values of the categorical variable InterventionLocationName with more than 100 records after cleaning:

```
'newhaven' 'stamford' 'hartford' 'wallingford' 'danbury' 'manchester' 'windsor' 'newbritain'
'easthartford' 'fairfield' 'westhartford''enfield' 'torrington' 'westport' 'norwich' 'norwalk' 'ridgefield'
'newtown' 'westhaven' 'groton' 'bridgeport' 'waterbury' 'vernon''branford' 'milford' 'berlin'
'southington' 'oldsaybrook' 'greenwich''newington' 'naugatuck' 'waterford' 'hamden' 'cheshire'
'newcanaan''northhaven' 'glastonbury' 'farmington' 'killingly' 'wilton' 'rockyhill''wethersfield'
'trumbull' 'middletown' 'bristol' 'mansfield' 'ansonia''tolland' 'monroe' 'newlondon' 'seymour' 'darien'
'orange' 'meriden''plainville' 'madison' 'stratford' 'westbrook' 'bloomfield' 'willington'
'colchester' 'marlborough' 'brookfield' 'guilford' 'southwindsor''bethel' 'montville' 'windsorlocks'
'simsbury' 'plainfield' 'newmilford''clinton' 'derby' 'easthaven' 'cromwell' 'watertown'
'windham''southbury' 'stafford' 'willimantic' 'storrs' 'putnam' 'griswold''woodbridge' 'ellington'
'litchfield' 'middlefield' 'redding' 'union''ledyard' 'haddam' 'somers' 'woodbury' 'coventry' 'grotoncity'
'thomaston' 'stonington' 'harwinton' 'northstonington' 'terryville''shelton' 'eastwindsor' 'brooklyn'
'chester' 'easthaddam' 'essex' 'pawcatuck' 'oldlyme' 'lebanon' 'thompson' 'deepriver' 'preston'
'canton' 'columbia' 'middlebury' 'eastlyme' 'lisbon' 'beaconfalls' 'winchester''bolton' 'avon'
'northbranford' 'eastgranby' 'suffield' 'newhartford''granby' 'easton' 'washington' 'unionville'
'burlington' 'hebron''oxford' 'coscob' 'ashford' 'goshen' 'durham' 'salem' 'sandyhook'
'plantsville' 'mystic' 'newfairfield' 'riverside' 'bethany' 'easthampton''andover' 'moosup'
'barkhamsted' 'prospect' 'pomfret' 'oldmystic''oakville' 'weston' 'salisbury' 'northcanaan' 'wolcott'
'bozrah''bridgewater' 'byram' 'chaplin' 'sherman' 'hampton' 'plymouth' 'sprague''franklin' 'portland'
'colebrook' 'centralvillage' 'woodstock' 'roxbury''eastford' 'canterbury' 'winsted' 'morris' 'bethlehem'
'sharon' 'kent''killingworth' 'glenville' 'oldgreenwich' 'norfolk' "oldg'wch"'chickahominy' 'sterling'
'cornwall' 'southport' 'canaan' 'westsimsbury''warren' 'voluntown' 'scotland' 'yourcity'
'grotonlongpoint''tariffville' 'm' '106' 'wauregan' 'mashantucket' 'northford' 'lostacres''quakerhill'
'pemberwick' 'hartland' 'eastmainstreet' 'lyme''cityofgroton' 'eastmainst' 'f' 'galesferry' 'weatogue'
'mainst''milldale'
```

Number of occurrences for each unique value of the categorical variable StatuteReason:

```
Speed Related             625023
Defective Lights         208260
Registration            207988
Cell Phone              200607
Moving Violation         176237
Other                  164193
Traffic Control Signal    162939
Stop Sign               150677
STC Violation            113289
Seatbelt                77053
Display of Plates         62391
Other/Error             43661
Window Tint              25287
Administrative Offense    24709
```

```
Suspended License        11519
Unlicensed Operation      8342
Equipment Violation       6174
Stop Sign                   31
```

Top 5 values with more records for the categorical variable ReportingOfficerIdentificationID:

```
WCW0264       8183
790642042     7765
1051          7166
1000002598    5562
1000002029    5430
```

Number of unique values for categorical variable ReportingOfficerIdentificationID: 8593
Top 5 values with more records for the variable InterventionDateTime:

```
03/07/2014 12:00:00 AM    269
03/08/2014 12:00:00 AM    266
03/17/2014 12:00:00 AM    260
03/21/2014 12:00:00 AM    252
03/22/2014 12:00:00 AM    250
```

Business questions technical support

Q: What is the target?
A: In a machine learning perspective, target is the variable that we wan't to predict.

Q: What are features?
A:In a machine learning perspective, features are the variables that we use to predict the target.

Q: What is a true positive?
A: In a prediction, a true positive is a value that was predicted as positive and its real value is positive, in conclusion, the prediction for positive is correct. In our context is when a vehicle is searched for contraband and contraband indices are found so the prediction for contraband (positive) is correct.

Q: What is a false positive?
A: In a prediction, a false positive is a value that was predicted as positive and its real value is negative, in conclusion, the prediction for negative is incorrect. In our context is when a vehicle is searched for contraband and contraband indices are not found so the prediction for contraband (positive) is incorrect.

Q: What is a true negative?
A: In a prediction, a true negative is a value that was predicted as negative and its real value is negative, in conclusion, the prediction for negative is correct. In our context is when a vehicle is not searched for contraband and contraband indices are actually not present in the vehicle so the prediction for not having contraband indices (negative) is correct. In our case,

in real scenarios, this cases are not possible to find out because if the vehicle was not searched is impossible to confirm of the vehicle has contraband indices or not.

Q: What is a false negative?
A: In a prediction, a false negative is a value that was predicted as negative and its real value is positive, in conclusion, the prediction for negative is incorrect. In our context is when a vehicle is not searched for contraband and contraband indices are actually present in the vehicle so the prediction for not having contraband indices (negative) is incorrect. In our case, in real scenarios, this cases are not possible to find out because if the vehicle was not searched is impossible to confirm of the vehicle has contraband indices or not.

Q: What is Amplitude?
A:In a group of numbers, amplitude is the difference between the maximum and the minimum of the group of numbers.

Q: What do you mean with people's background?
A: With people's background we mean people's age, gender, sex, race…

Q: What do you mean with protected groups?
A: protected groups are people with the same age range, gender, sex and race. There are as much protected groups as all possible combinations of these referred variables.

Model technical analysis

Age: From numeric variable to categorical variable

Age	Value
<18	Youth
>=18 and <=35	Young Adult
>=36 and <=55	Adult
>=56	Senior

Table 4: Age conversion from numeric variable to categorical variable.

Percentage of Positives Encoding

This encoding is used on features: InterventionLocationName, StatuteReason and InterventionReasonCode. For each feature that we apply this encoding, the dataset is grouped by the considered feature. For each group is calculated the percentage of contraband found per number of total stopped vehicles. With this we can construct a dictionary with unique values and its correspondent percentage of positives. This means that each unique value is replaced by the number of rows that have this unique value and the target is true, divided by the number of rows that have the considered unique value.

Model Selection: F1-Score Macro Average

	Undersampling	Oversampling	Smote
Random Forest	0.63	0.64	0.66
GradientBoosting	0.65	0.65	0.66
AdaBoost	0.70	0.70	0.71

Table 5: Models Results.

SearchAutorizationCode Regularization

	threshold	mean	std	f1 macro avg	n < 5
Prediction	20	6.57	7.19	0.71	289
Real	20	10.4	7.63	-	80

Table 5: Distribution of amplitude on precision for protected groups by department name and model score when regularization is not applied to SerachAutorizationCode.

(threshold, power)	mean	std	f1 macro avg	n < 5
(0,1)	2.79	2.9	0.45	5025
(0,2)	2.76	2.94	0.45	5203
(20,1)	7.28	7.26	0.71	228
(20,2)	7.25	7.22	0.71	232
(70,1)	6.79	7.01	0.70	269
(70,2)	7.00	7.30	0.71	253

Table 6: Distribution of amplitude on precision for protected groups by department name and model score when amplitude regularization is applied to SerachAutorizationCode

(threshold, power)	mean	std	f1 macro avg	n < 5
(0,1)	7.41	7.24	0.71	223
(0,2)	7.42	7.27	0.71	221
(20,1)	7.99	6.70	0.68	227
(20,2)	8.11	6.93	0.67	247
(70,1)	5.85	5.72	0.55	972
(70,2)	6.10	5.77	0.56	824

Table 7: Distribution of amplitude on precision for protected groups by department name and model score when precision regularization is applied to SerachAutorizationCode

threshold	mean	std	f1 macro avg	n < 5
20	2.65	2.86	0.44	5457

Table 8: Distribution of amplitude on precision for protected groups by department name and model score when SerachAutorizationCode variable is not included in the model.

Roc Curve - Chosen Model

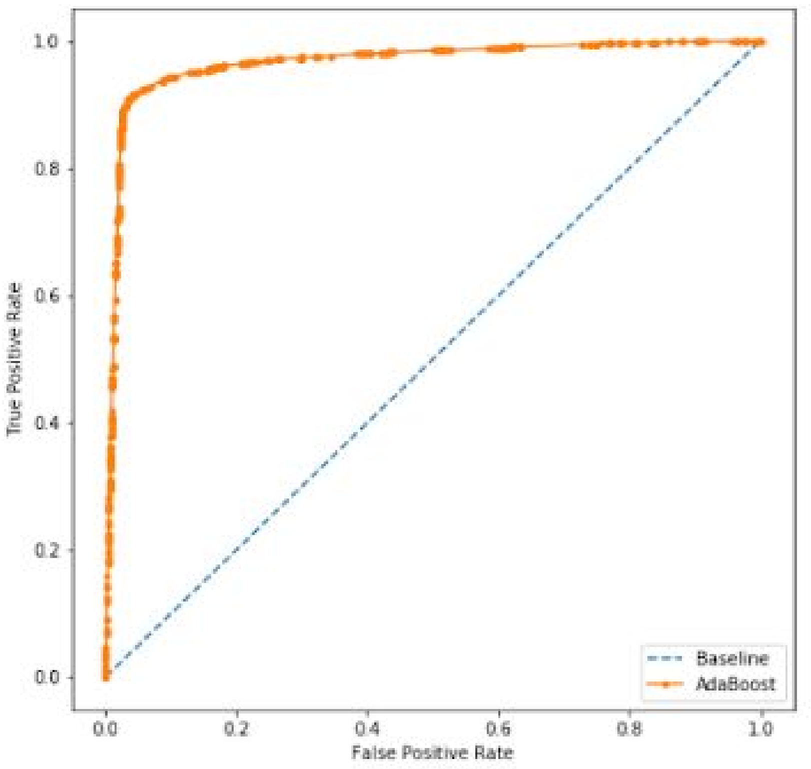


Table 17: ROC Curve for the chosen model. Area under roc curve is 0.97.