Client Requirements

Summary

The goal of this project is to analyze and improve the stop and search policy of the United Kingdom Police Department, that's supposed to ensure that officers stop people and cars when there is probable cause. There are 41 police stations in the United Kingdom, each with its own policy. There have been accusations in the press that the police tend to stop and search certain minorities at a higher rate than others and that during searches women of certain age groups and ethnicities are asked to remove cloth more frequently than other groups.

Therefore the first task of the project is to search for evidence that any police station may be discriminating on gender, ethnicity or age regarding who they chose to stop, and also who they ask for any clothing to be removed.

If the previous task is successful, the second task of the project is to create and deploy an application for authorizing searches. This application should ensure uniformity of stop and search policies. It will be integrated on the police system that officers need to use and will be running during 2020. A search should be authorized only when there is likelihood above 10% for the search to be successful. This requirement will be evaluated per station and search objective. Furthermore it is also expected that the application should be able to level the search success rate between ethnicities for every station and for every search objective, without significantly diminishing the current ability to detect offences.

Requirements clarifications

We requested Dr. Wilson to clarify some of the requirements which we found ambiguous. The most sensitive requirement of this project is related to non-discrimination of populations groups. Previous reports on this subject defined different metrics for the discrimination. One is the disproportion in stop and search between groups, measured as the number stop and searches per the total population for the group. Moveover the population can be normalized by the resident population, which can be obtained from Census data (not used in this analysis). Another is the disparity in stop and search success rate, measured as the number of successful searches per the total number searches for the group. Dr. Wilson clarified that we should analyze the search success rate, as she expected some disproportion due to correlations with other factors such as economic status.

Furthermore, a search is considered successful when the outcome is different from 'Nothing found - no further action' and 'A no further action disposal' and it is linked to the object of search. When looking for evidence of discrimination we should divide the population in groups with a given ethnicity and gender for each police station. The difference in the search success rate between groups should not be higher than 5%. The difference in average search success rate

between stations should not be higher than 10%. A group with less than 30 observations is considered not statistically significant.

Regarding the analysis on the policy for asking cloth removal, for each station the population should be divided in groups with a given gender, ethnicity and age. The same discrimination criteria as mentioned above will be used.

The project should deliver an API to an application used to approve or reject searches. The performance of the API will be evaluated using two metrics : the probability that an offence is found in an authorized search (precision) and the probability that all offences are present in the set of authorized searches (recall).

$$Precision = \frac{\#\ Authorized\ searches\ with\ positive\ outcome}{\#\ Authorized\ searches\ with\ positive\ outcome + Authorized\ searches\ with\ negative\ outcome} > 10\%$$

$$Recall = \frac{\#\ Authorized\ searches\ with\ positive\ outcome}{\#\ Authorized\ searches\ with\ positive\ outcome + \#Non\ Authorized\ searches\ with\ positive\ outcome} > 90\%$$

The threshold for the first metrics was an explicit requirement. For the second metric we chose a threshold that would result in a low rate of missed offences in non-authorized searches.

The difference in the search precision between different ethnicities groups for a given station and object of search should not be higher than 5%. The difference in the average precision between stations should not be higher than 10%.

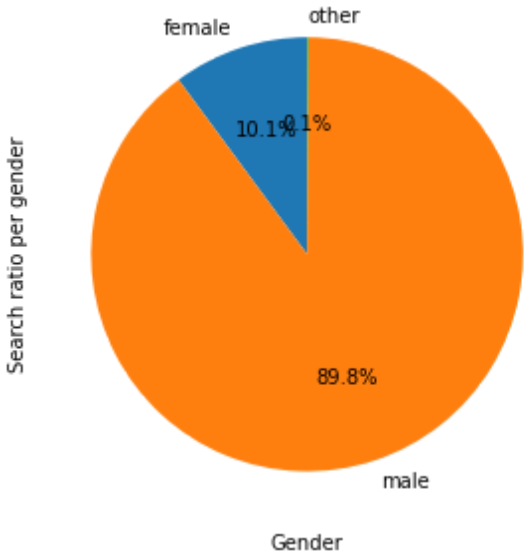
Dataset Analysis

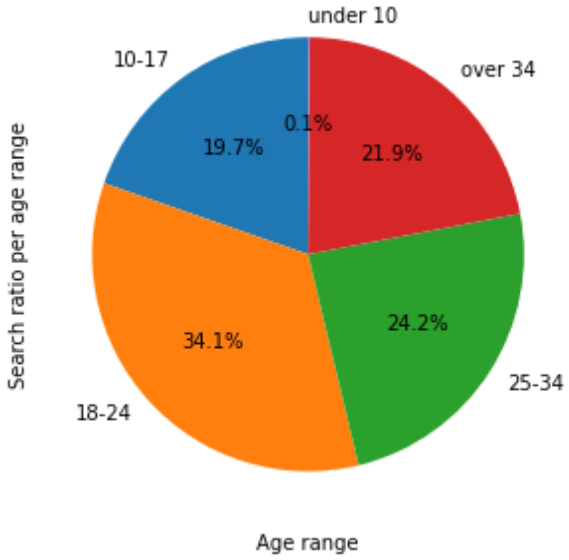
General Analysis

The dataset for stop and police search spans the period between December 2017 and December 2019. In total 660611 stop and search operations were registered. More than 50 % were performed by the Metropolitan police station. For this station there is no information on whether the outcome is linked to the object of search and if cloth was asked to be removed. Therefore we exclude this station from the analysis. Of the remaining 41 stations there are some which search more than others, like MerseySide and Essex (see Figure 2 of Annex “Dataset technical analysis “), probably because of the different size of population and number of officers.

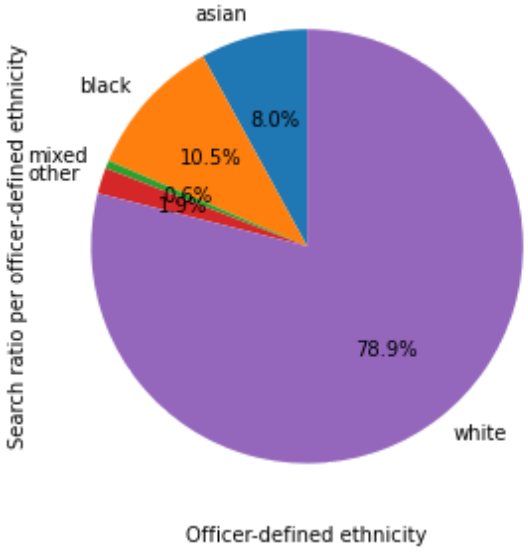
In this dataset there is an imbalance in the proportion of stop and searches between population groups. Only 10% of the searches are performed on females. Most of the searches are done for

young ages (only 22% of them refer to the population above 24 years old). Surprisingly there are entries for very young kids, with age below the criminal responsibility (10 years).

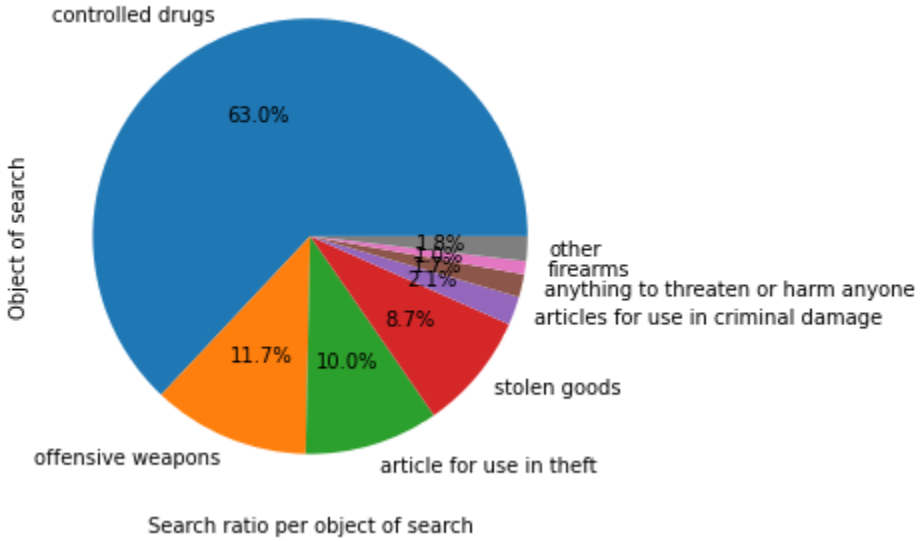




The ethnicity of the subject being searched is defined by the officer conducting the search. Most of the searches were performed on white subjects (80%) (see figure below) . There is also data on the subject self defined ethnicity. The ethnicity is most of the time correct when the subject self-defined ethnicity is white, black and asian. When subjects are mixed or of other ethnicity the officer defined ethnicity is incorrect most of the time. For numeric details see Figure 1 of Annex “Dataset technical analysis”.

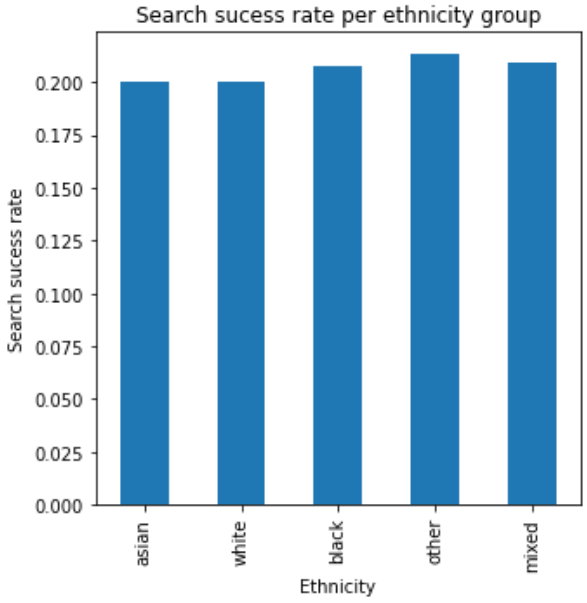


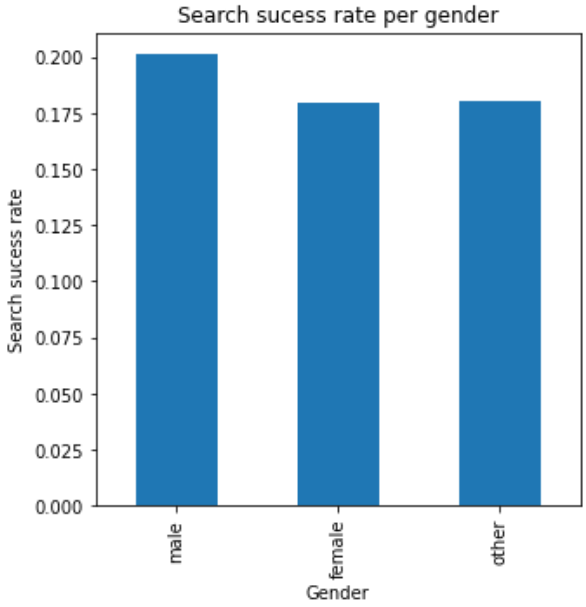
Below we present the proportion of search per type of object being searched. It is also imbalanced. 60% of the searches correspond to “controlled drugs”. “offensive weapons”, “articles for use in theft” and “stolen goods” are also searched frequently. Other objects are seldom searched.



Several imbalances were found for the proportion of searches per group. If there is no disparity in the search success rate between groups, these imbalances don’t provide enough support for presence of discrimination.

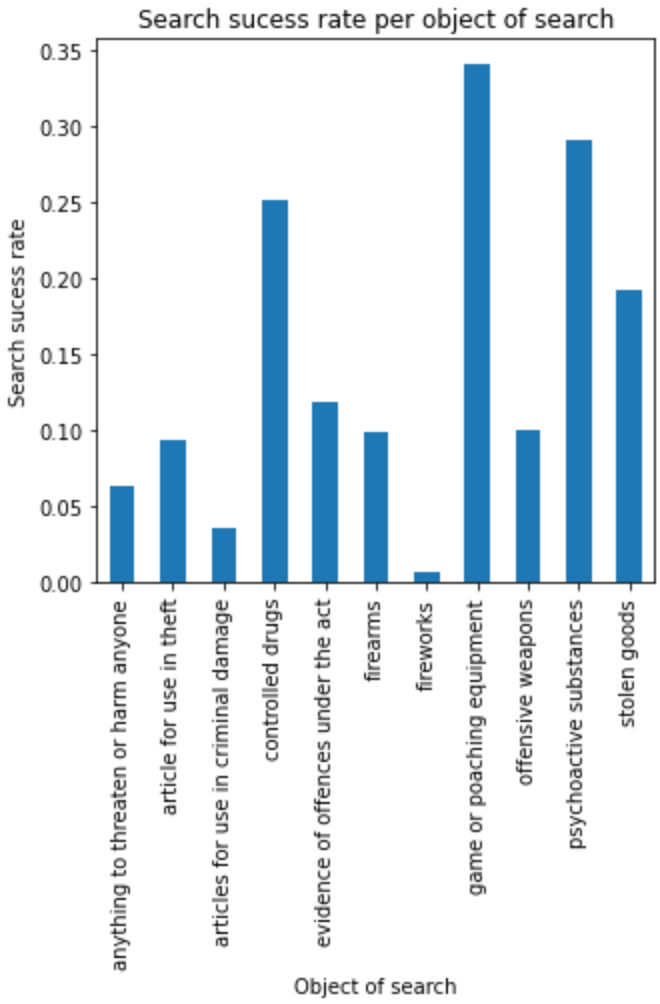
The overall search success rate is approximately 20%. Below we plot the search success rate per ethnicity and gender groups.



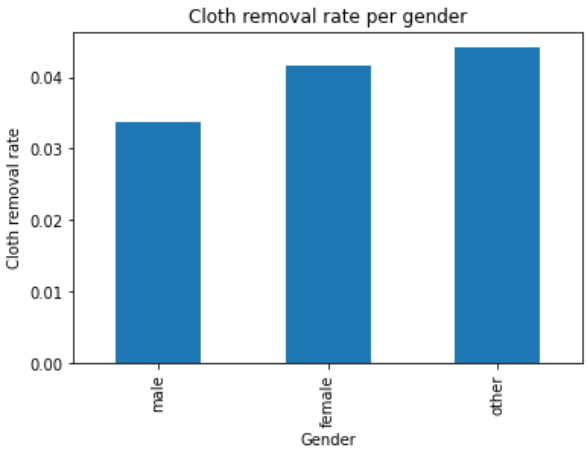


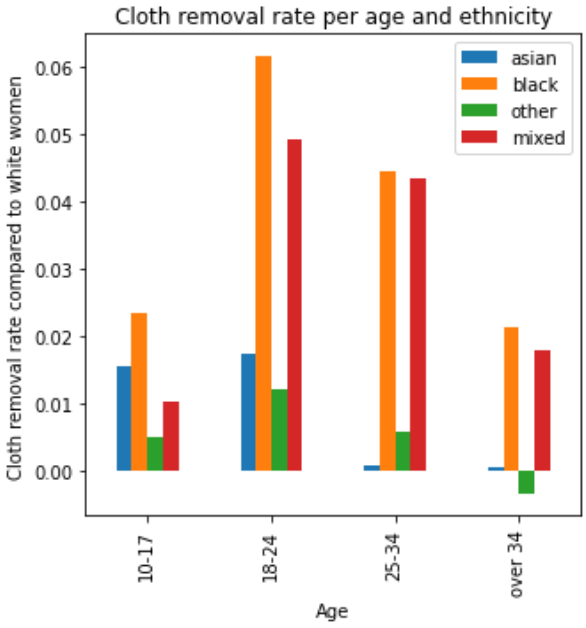
The disparity in the search success rate for gender and ethnicity groups is below 5%. Therefore there is insufficient statistical evidence for discrimation in the aggregate of all the stations.

There is disparity however in the search success rate per object of search. Searches for “Controlled drugs” yield a success rate a bit above the overall rate, whereas “offensive weapons”, “firearms” and “articles for use in theft are half of the overall rate”, which mean they are being over searched. On the other hand searches for “wildlife offences”, “crossbows” and “goods on which duty have not been paid” have no statistical significance.



Regarding the analysis on the policy for asking cloth removal during searches, there is no significant disproportion between gender groups, as can be seen in the plots below. However if the rate of asking for cloth removal is compared between women of the same age and different ethnicity, there is an imbalance. “Black” and “Mixed” women are asked to remove clothes more frequently than white women of the same age. The disproportion is bigger for ages between 18 and 34 years, nevertheless being smaller than 6%.





Business questions analysis

The previous analysis was performed for the aggregate of all police stations. To answer the business questions we also need to evaluate if there is disparity among individual stations.
We calculated the disparity between ethnicities and genders for all the police stations regarding search success rate, measured as:

$$search\ success\ rate\ = \frac{\#\ searches\ in\ group\ X\ with\ positive\ outcome}{\#\ searches\ in\ group\ X}$$

The plot belows are scatter plots where each point represents the disparity between white and non-white groups (left plot) and male and female groups (right plot) for a station. If there was no discrimination at all, we would expect all the points to sit on the diagonal line. This is not the case.
There are stations where the disparity between ethnicities is above 5%, which is the threshold specified by the client in the clarification of the requirements . The success rate for non-white groups is significantly lower for searches in the City of London, Bedfordshire, Btp, Nottinghamshire and 'Suffolk'. (See Annex “Business question technical support). When the officer defined ethnicity is “mixed” or “other” the disparity is greater than when it is “black” or “asian”. In West Mercia and Cambridgeshire non-white have higher search sucess rate than white.
For most of the stations the search success rate is higher for men than for women, which might imply that the latter are being over-searched.

Figure 1 : Search success rate per station and ethnicity.

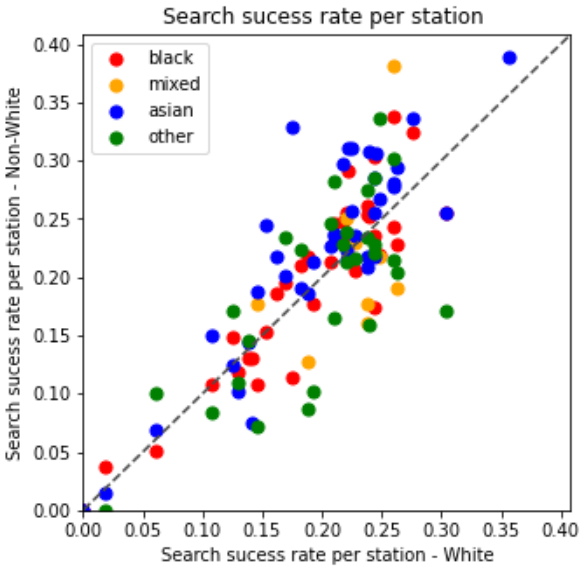
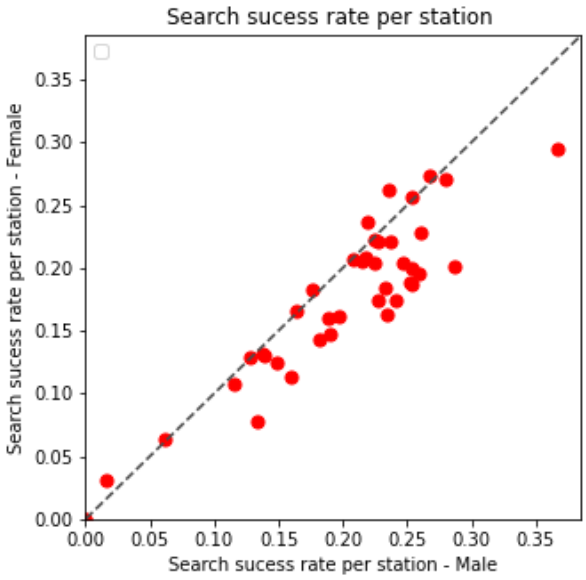


Figure 2 : Search success rate per station and gender



The difference in average search rate is also sometimes higher than 10% for some stations. If all stations had the same search rate and did not discriminate, we would expect all the points of the scatter plot to be concentrated in the same region and on the diagonal line.

The search success rate presents also a disparity when comparing the type of object being searched. Psychoactive substances, offensive weapons and firearms have lowers sucess rates for non-white than for whites. Controlled drugs, psychoactive substances and firearms have lower success rates for females than for males.

Figure 3: Search success rate per object and ethnicity

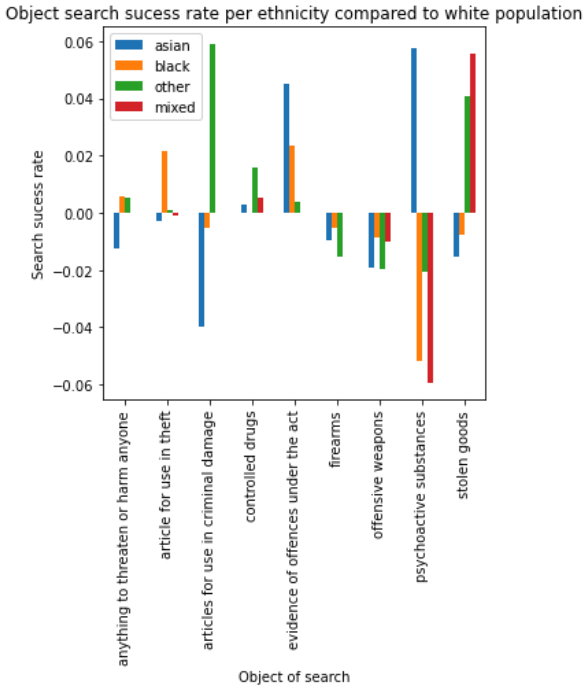
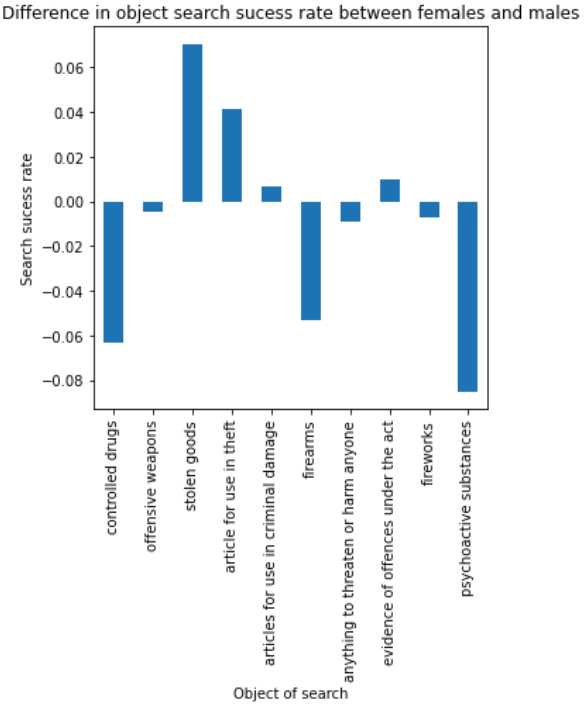


Figure 4: Search success rate per object and gender



Regarding the policy on asking for cloth removals, we find no significant evidence of discrimination between etnicities when both genders are grouped. The difference in search success rate is also computed by age group and compared to the category with the most senior people “over 34 years” . It is higher for younger age groups.

Figure 5: Search success rate per ethnicity when clothes are removed

Figure 6: Search success rate per age range when clothes are removed

The same analysis was restricted to female groups according to the age group. The search success rate tends to be higher when officers ask to remove cloth for younger females.

Figure 7: Search success rate per age range for female women when clothes are removed.

We don’t have enough statistics to split females according to ethnicities.

Conclusions and Recommendations

From the analysis of current dataset we conclude that:

- There is discrimination between ethnicites for 23 out of the 41 stations (absolute difference between search success rate is greater than 5%).
- In 12 of the 23 stations which discriminate, the disparity occurs only for “mixed” and “other ethnicities”.
- In 11 of the 23 stations which discriminate, the search success rate is lower for non-white than white.
- The search success rate is lower for women than for men.
- The difference in average search success rate is higher than 10% for some stations.
- There is not enough statistical evidence of ethnicity or age discrimination when asking to remove clothes.

Table 1: Discrimination metric in the training dataset

# Stations with discrimination	# Stations with positive discrimination	# Stations with negative discrimination	Median of total discrimination	Spread of total discrimination
23	14	11	4.9%	49.0%

The policy on searches deserves further investigation regarding discrimination. All groups should have similar search success rates. We recommend that the training policies regarding discrimination are reviewed for the stations with higher disparities. The list of stations for which there is evidence of discrimination can be consulted in Annex “Business questions technical support”.

Modeling

Model expected outcomes overview

We have developed a Machine Learning model which will be deployed in an application and shall be used to approve or reject searches during the year of 2020. It is expected to satisfy the requirement of authorizing searches only when there is a likelihood above 10% for the search to be sucessful. Only a very small fraction of offences (< 10%) should be missed due to non authorized searches. This expectation was confirmed on a validation dataset.

To attempt to reduce the discrimation between ethnic groups for some stations, the model does not use directly the projected subject characteristics : ethnicity, gender and age. The model is also agnostic to the station of the search and uses a single decision boundary. However when applying the model on a validation set the discrimination was not significantly reduced. Neither

the difference in the average search success rate between stations and object of searches was lower than 10%. Therefore these requirements are not expected to be satisfied. Since there disparity in search success rate between etchnic groups depends on the station and the object of search, it is not sufficient to suppress the protected characteristic when training the model. Despite the station not directly used in the model, the location of the search (latitude and longitude) is correlated to the area of the station. Creating discrimination free models trained on data with discrimination is a non trivial problem. We tried to solve it by calculating the number of samples for each subgroup in the training dataset for which the success rate would be constant and appling resampling with substitution for the training dataset. This approach was not successful at reduction the discrimination. Therefore there should be more research on data preprocessing techniques which can compensate for the imbalances found in this training dataset.

The client will receive addresses of two endpoints for API: one for authorizing the searches and another to provide the true outcome of the search. From our tests we expect that the API will respond with a decision in less than 0.5 seconds (latency).

Model specifications

The model should either approve or reject searches based on the expected outcome of the search (successful/unsuccessful). Since this is a classification problem we choose to model it with a Random Forest Tree classifier. This is a ensemble model, used to avoid overfitting of a single tree model to dataset.

The dataset was further split in training and validation subsets, each having 70% and 30% of the dataset respectively. Since there is an imbalance in the target of the model, the splitting was stratified according to the proportion of successful searches.

The categorial features of the dataset were converted to numeric using One-Hot encoding. Missing values in the categorical features “Part of a policing operation” and “Legislation” were handled as a separate category. The only numerical features with missing values are “Latitude” and “Longitude”. These missing values were mapped to zero in order to create a separate category which does not overlap with the range of latitudes and longitudes of the dataset. The numeric values were scaled by removing the mean and scaling to unit variance.

To prevent potential discrimination the protected features “Age range”, “Officer-defined etchinicty” and “Gender” were dropped from the training dataset. Features “Outcome”, “Outcome linked to object of search”, “Self-defined ethnicity” and “Removal of more than just outer clothing” are not available before the search is approved, so they were also dropped from the training set. The features “Year” and “Minute” were dropped due to lack of sensitivity.

Since there are subgroups of the population for which there is not enough statistics, we require that a leaf of a tree should have at least 30 samples. The trees should also be balanced

according to the target distribution. More details on model implementation and parameters can be found in Annex “Model technical analysis”.

Once the model is trained we test it using the validation dataset. The performance of the model is measured with two metrics: precision and recall. The precision evaluates if the approved searches are likely to be successful. The recall evaluates if the approved searches are likely to discover all the criminal offences, that is, if in rejected searches there are no offenses. We require the precision to be at least 20% (the current average search success rate). The recall should be maximized. We consider that there is enough evidence of discrimination if the difference between the precision (or search success rate) for different groups is greather than 5%. The difference between the average precision for different stations should not be larger than 10%. When the subgroup has less than 30 samples, we don’t evaluate discrimination. The probability threshold set to authorize search was 0.33, because it led to a recall above 90%.

Analysis of expected outcome

Applying the model to the validation set we get the following estimates for the performance metrics:

- Search success rate (precision) = 23%.
- Discovery rate (recall) = 93%.
- ROC score = 0.66.

The model has a low precision but a high recall. It prefers to authorize a search which would be unsuccessful than to reject a search than would be successful. Since the probability of a search to be successful is higher than 10% and the fraction of non-authorized searches which would correspond to a true outcome is small, the client requirements for the performance metrics are satisfied. However since the ROC score is only 66% higher than for a random classifier, the model is not very good at predicting if a search will be successful or not.

The importance of the features for the model was obtained for sklearn classifier attribute ‘feature_importances_’ :

1. Object_of_search = controlled drugs (0.161754)
2. Latitude (0.129942)
3. Longitude (0.095683)
4. Legislation =misuse of drugs act 1971 (section 23) (0.086523)
5. Legislation = police and criminal evidence act 1984 (0.077897)
6. Hour (0.059819)
7. Day (0.048423)
8. Object of search =stolen goods (0.042743)
9. Month (0.041158)
10. Object of search = offensive weapons (0.040847)

11. Object of search = article for use in theft (0.038385)

The most important features are the object of search and the coordinates of the search. In section "business questions analysis" it was shown that the search success rate per object depends on the ethnicity. Therefore it is expected that model will not be able to remove discrimination between ethnicities.

We define the metric for the discrimination as the difference in precision between a non-white group (black, asian, other or mixed) and the white group.

$$discrimination_{\,non\,white\,group} = precision_{\,non\,white\,group} - precision_{\,white\,group}$$

In Annex "Model Technical Analysis" , section "Discrimination metrics", it is plotted the discriminination metric per station for the validation set. We can conclude that

- There is no evidence of discrimination between ethnic groups for 24 out of the 41 police stations.
- There is evidence of discrimination (above 5% disparity in search success rate) between ethnic groups for 17 out of the 41 police stations .
- In 11 of the 17 stations which discriminate, the disparity occurs only for "mixed" and "other ethnicities".
- In 9 of the 17 stations which discriminate, the search success rate is lower for non-white than white.

We summarize this information in the table below, where we also compute the median and spread of the total discrimination per station:

$$total\ discrimination_{\,station\ i} = \sum_{all\ non\ white\ groups,\ station\ i} discrimination_{\,non\,white\,group}$$

Table 2 : Discrimination metrics of the model for validation set

# Stations with discrimination	# Stations with positive discrimination	# Stations with negative discrimination	Median of total discrimination	Spread of total discrimination
17	11	9	4.7%	48.0%

Therefore the client requirements linked to non-discrimination of populations groups are not satisfied. There was no significant improvement in non-discriminaton metrics with respect to the current policy.

Alternatives considered

Since the training dataset is imbalanced we tried to apply resampling. Groups with lower search success rate should have less samples than in the current dataset and the ones with higher success rate should have more. Since there are groups which have low statistics we used sample replacing. This approach seemed promising but did not decrease the discrimination metrics in the validation dataset, either because the resampling strategy was incorrect or due to model overfiting to training dataset.

We also considered other Machine Learning models : Logistic Regression and Gradient Boosting Classifiers. None was able to diminish significantly the metrics for discrimination.
Since the training data covers at least two calendar years, we trained a RandomForest model in the most recent year (2019) because there might have been training towards non-bias in officer searches. The results were unchanged.

When analysing the dataset we already found that the officer defined ethnicity is mostly wrong for “mixed” and “other ethnicities". Therefore we tried to assign different ethnicities to those groups in the training dataset , e.g considering that “mixed” subjects should be identified by the officer as “black” subjects and subjects of “other” ethnicities should be identified as “white”. This approach was not sucessful at reducing discrimination.

Known issues and risks

The current model does not use protected characterics which are a potential source of discrimination to approve the searches. Though it was not able to reduce significantly the discrimination found in the training dataset. This a significant limitation and risk which needs further investigation.

The hyper parameters of the Random Forest model were not tuned, such as number of trees and max_depth of each tree. Most of the effort was spent trying different approaches to reduce discrimination of the model. Therefore there is room for optimization of the precision and recall of the model.

On the other hand it uses features such as coordinates and date of the search to make decisions. If during production there is a change in pattern of criminality correlated to these features, the model might provide different precision and recall metrics. If this happens it might need to be retrained.

If during production the features have different distributions over the target class (outcome true and linked to object of search) than the ones used to train the model the performance of the model will decrease with regard to the expectations. If there are temporary changes in the production data , e.g. lockdowns are in place during certain periods due to COVID19 and the number of searches is decreased or some crimes become more frequent than others, a new feature must be added (e.g. lockdown in place : true or false) and the model needs to be retrained to predict differently depending on the value of the feature. If the changes are seasonal, e.g. crime rate increases during winter months due to unemployment, a seasonal feature must be added and the model needs to be retrained. If the changes are permanent the model needs to be retrained excluding previous data. In every scenario data must be continuously monitored to detect changes in features distribution, model predictions and model probability.

When unknown values are imputed for the categorical features “Type”, “Legislation” and “Object of Search”, the deployed model will replace them by the default values “Person Search”, “Misuse of drugs act 1971 (section 23)” and “controlled drugs” respectively. This choice might lead to bias in the predictions. It is recommended to retrain the model if unknown values are detected.

In case extra features are added or existing features are removed from the API request input specification , the model will return error messages.

Model deployment

Deployment specifications

The API URL for the model deployed is https://ldssa-batch4-cap-pribeiro80.herokuapp.com/.
It provides two endpoints : **/should_search/ and /search_result/**.

1) /should_search/
   This endpoint expects to receive a request with the following content

```
{
 "observation_id": <string>,
 "Type": <string>,
 "Date": <string>,
 "Part of a policing operation": <boolean>,
 "Latitude": <float>,
 "Longitude": <float>,
 "Gender": <string>,
 "Age range": <string>,
```

"Officer-defined ethnicity": <string>,
"Legislation": <string>,
"Object of search": <string>,
"station": <string>
}

It returns an answer with the observation_id and the predicted outcome (whether the officer should search or not).

{"observation_id": <string>,
"outcome": <boolean>}

If the received observation ID already exists on the database, it returns HTTP code 405 with the error message "Incorrect or missing data."

2) /search_result /
This endpoint expects to receive a request with the following content

{"observation_id": <string>,
"outcome": <boolean>}

It returns an answer with the observation ID, the predicted outcome and the true outcome given:
{"observation_id": <string>,
"predicted_outcome": <boolean>,
"outcome": <boolean>
}

If the received observation ID does not exist on the database, it returns HTTP code 405 with the error message "Invalid ID supplied".

In order to reproduce the deployment of the model the following workflow should be used :

1. Train the model using the training dataset.
2. Serialize the model, the name of the features that it uses and its data types.
3. Create a application which:
    a. Unserializes the previous objects
    b. Validates the parameters of the request containing the search to be approved.
    c. Use the model to predict the success of a search based on its parameters.
    d. If the model predicts the search to be successful with probability greater than 20% , approve the search, otherwise reject it. Send an answer with this result.
    e. Save the parameters of the search to a Database, to allow monitoring and inspecting of the results.
    f. Allow the user to record if the search was successful or not.

4. Deploy the application using a cloud service.

Our model was coded in python and depended on pandas, numpy and sklearn libraries. An archive containing the source code will be sent later. It was deployed in an application with two endpoints: one to predict if search is successful and other to update the result of the search.

The application was deployed to Heroku and used the PostGres database.

Known issues and risks

We have tested that the application does a proper validation of the inputs and can handle missing values in the parameters of the search. Nevertheless there is a possibility of application malfunction in these scenarios.

During our tests the API of the application was able to serve requests in less than 0.5 seconds. However we did not perform extensive load tests. If during 2020 the rate of searches per minute is very high for the whole UK, the response time might be higher. We cannot ensure a maximum value for the API response time during periods of intensive load.

Annexes

For search success rate samples with less than 30 were removed

Dataset technical analysis

We performed the following transformations on the dataset

- To define the target feature of the analysis we created a new boolean feature which is true if and only if the feature “Outcome” is different from 'Nothing found - no further action' and 'A no further action disposal' and feature “Outcome linked to object of search” is true”.
- Missing values for the feature “Outcome linked to object of search” were assumed to correspond to “False”.
- Missing values for feature “removal of more than just outer clothing” were assumed to correspond to “False” for non-vehicle searches.
- All rows of the dataset provided by the customer which corresponded to Metropolitan police station were removed.
- From the values of the feature ‘Date’ we created 5 features: “Year”, “Month”, “Day”, “Day of week”,”Hour”,”Minute”.
- To analyze if the self-defined ethnicity matched the officer defined ethnicity we defined an intermediate feature “etchinity” which aggregates different ethnical groups reported by the subjects into larger groups corresponding to the officer self-defined ethnicity.

We performed a cross tabulation of both sources of ethnicity (officer and self-defined), normalized by the subject self defined ethnicity. If the officer was able to correctly identify the ethnicity, the matrix on the left plot would have 1's in the diagonal entries and 0's in the remainder.

Figure 1 : Confusion matrix for subject ethnicity

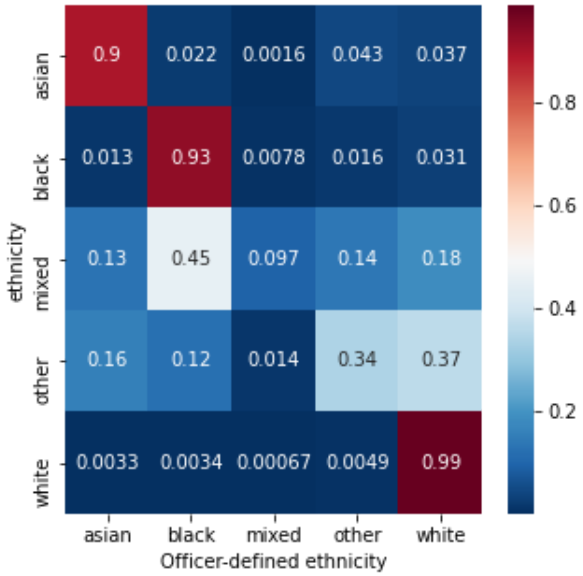
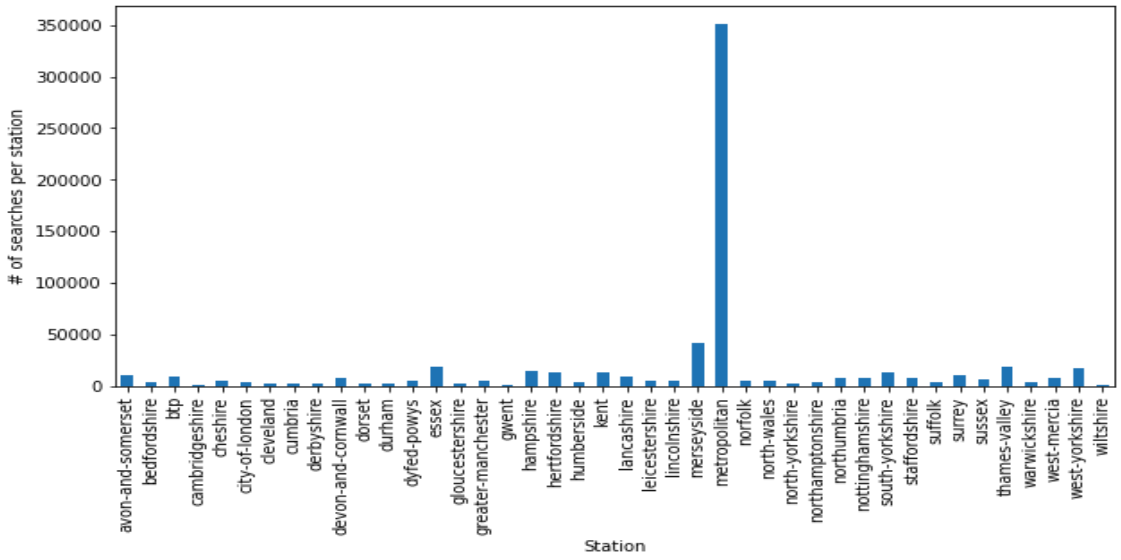
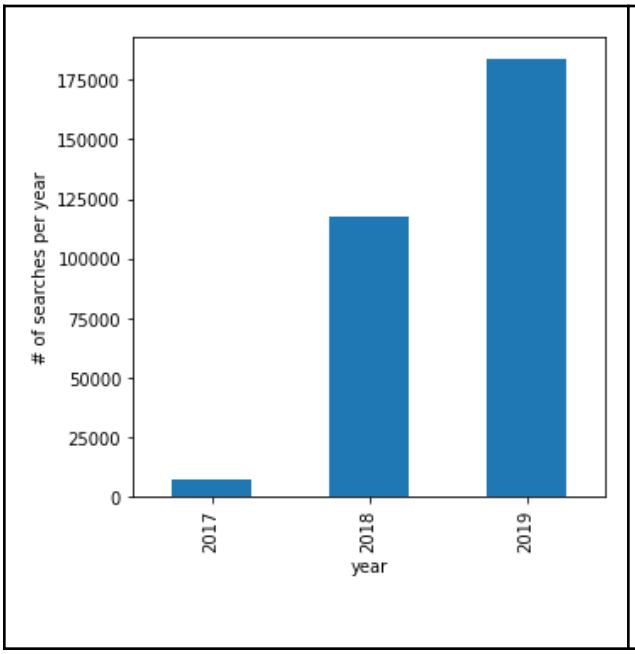
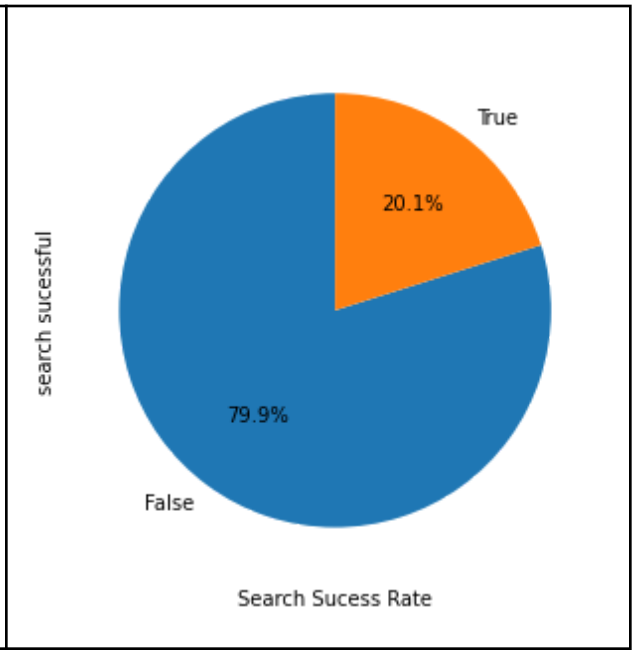
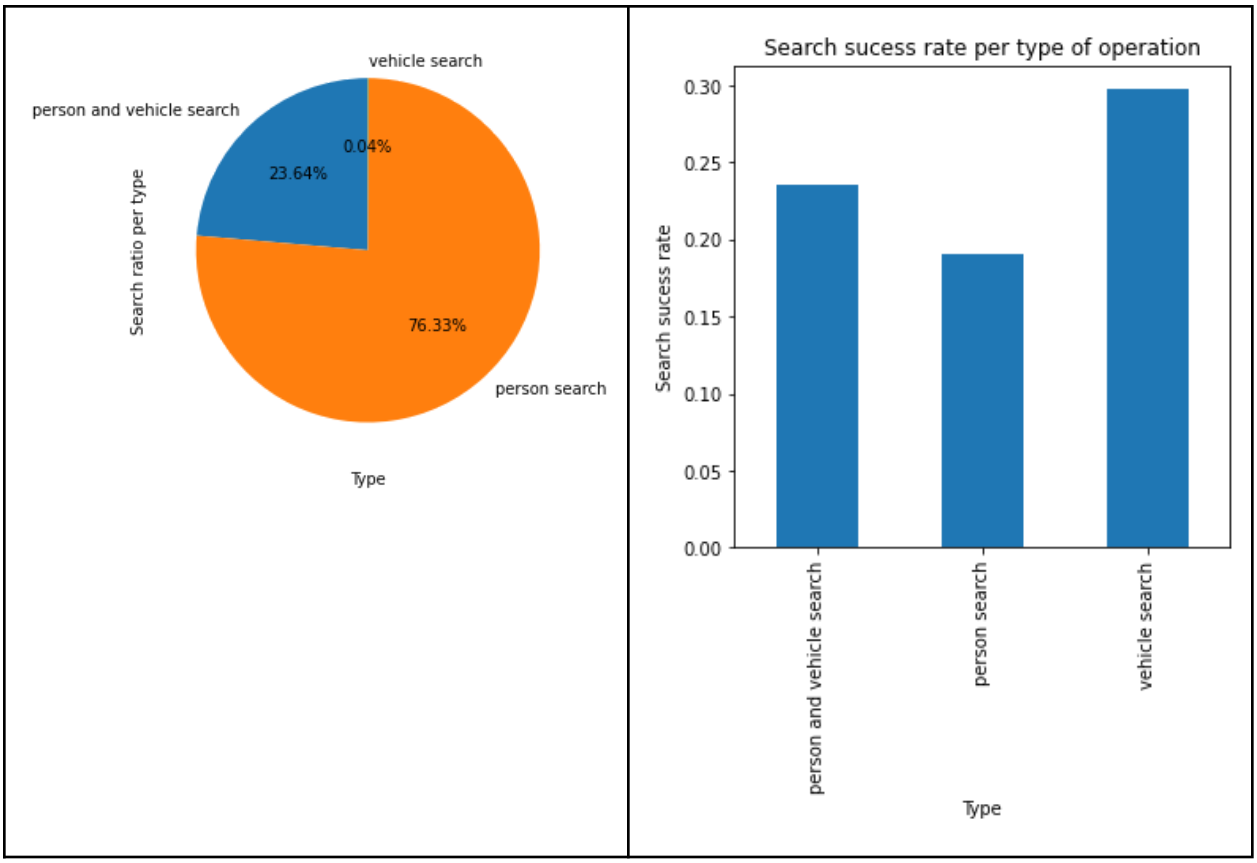


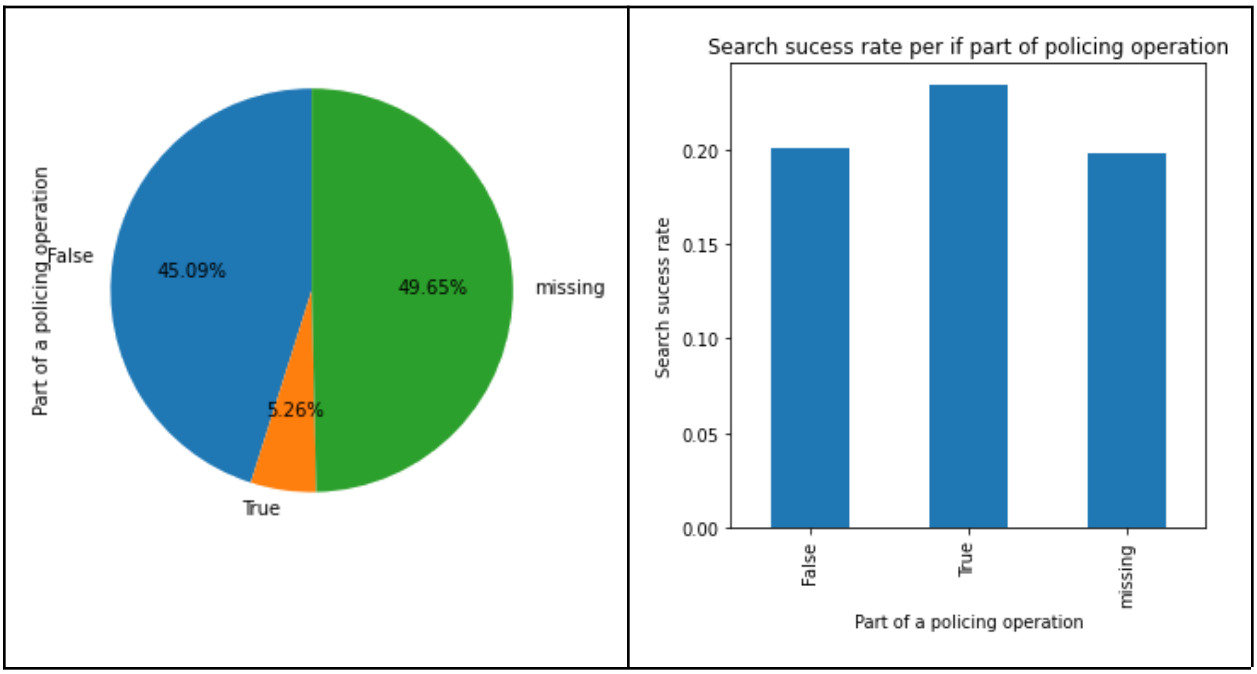
Figure 2 : Number of searches per station


# of searches per station
350000
300000
250000
200000
150000
100000
50000
0
avon-and-somerset
bedfordshire
btp
cambridgeshire
cheshire
city-of-london
cleveland
cumbria
derbyshire
devon-and-cornwall
dorset
durham
dyfed-powys
essex
gloucestershire
greater-manchester
gwent
hampshire
hertfordshire
humberside
kent
lancashire
leicestershire
lincolnshire
merseyside
metropolitan
norfolk
north-wales
north-yorkshire
northamptonshire
northumbria
nottinghamshire
south-yorkshire
staffordshire
suffolk
surrey
sussex
thames-valley
warwickshire
west-mercia
west-yorkshire
wiltshire
Station


# of searches per year
175000
150000
125000
100000
75000
50000
25000
0
2017
2018
2019
year


search sucessful
True
20.1%
79.9%
False
Search Sucess Rate

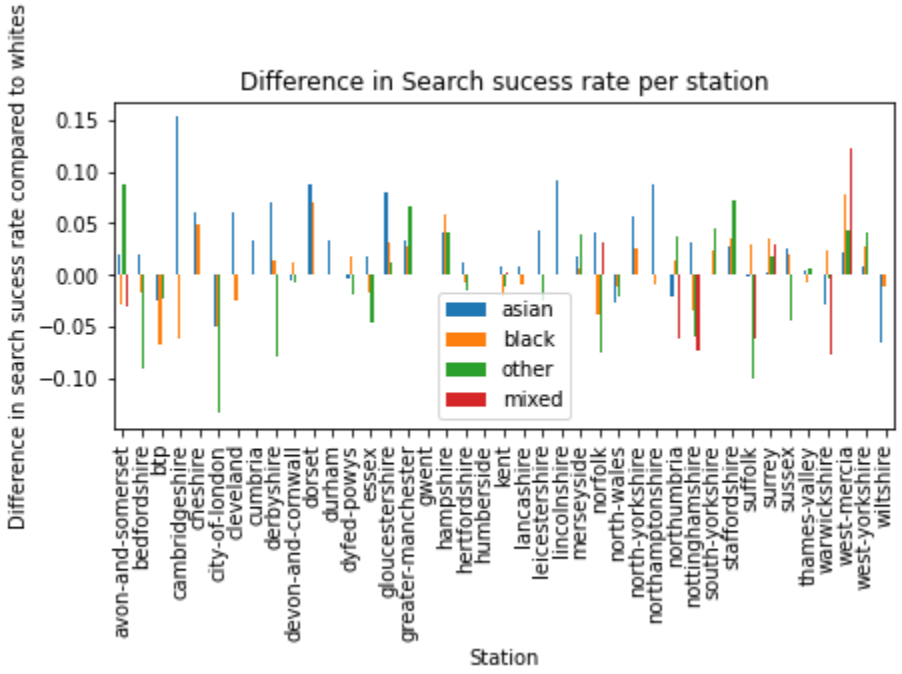

vehicle search
person and vehicle search
0.04%
23.64%
76.33%
person search
Search ratio per type
Type
Search sucess rate per type of operation
Search sucess rate
0.30
0.25
0.20
0.15
0.10
0.05
0.00
person and vehicle search
person search
vehicle search
Type


Part of a policing operation
False
45.09%
49.65%
missing
5.26%
True
Search sucess rate per if part of policing operation
Search sucess rate
0.20
0.15
0.10
0.05
0.00
False
True
missing
Part of a policing operation

Business questions technical support

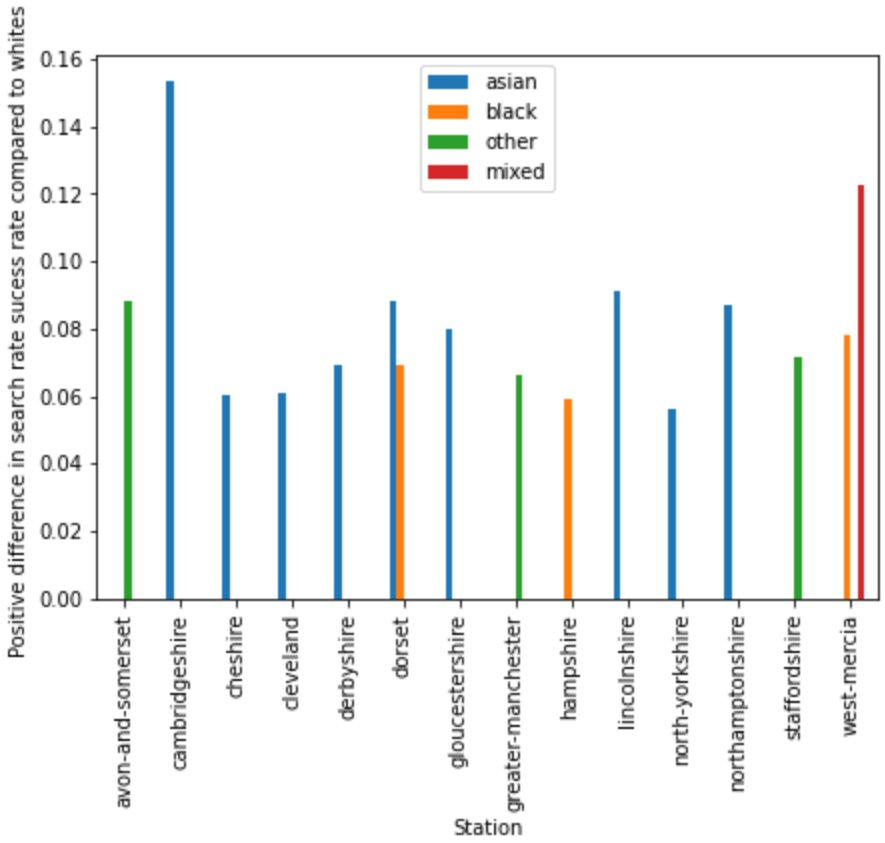
Stations for which absolute difference in the search success rate between non-white and white is greather than 5% :
‘Avon-and-Somerset', 'Bedfordshire', 'Btp', 'Cambridgeshire','Cheshire', 'City-of-london','Cleveland', 'Derbyshire', 'Dorset', 'Gloucestershire', 'Greater-Manchester', 'Hampshire', 'Lincolnshire', ‘'Norfolk', 'North-Yorkshire', 'Nort-Hamptonshire', 'Northumbria', 'Nottinghamshire', 'Staffordshire', 'Suffolk', 'Warwickshire', 'West-Mercia', 'Wiltshire'

Figure 1 : Difference in search success rate between non-whites and whites for each station



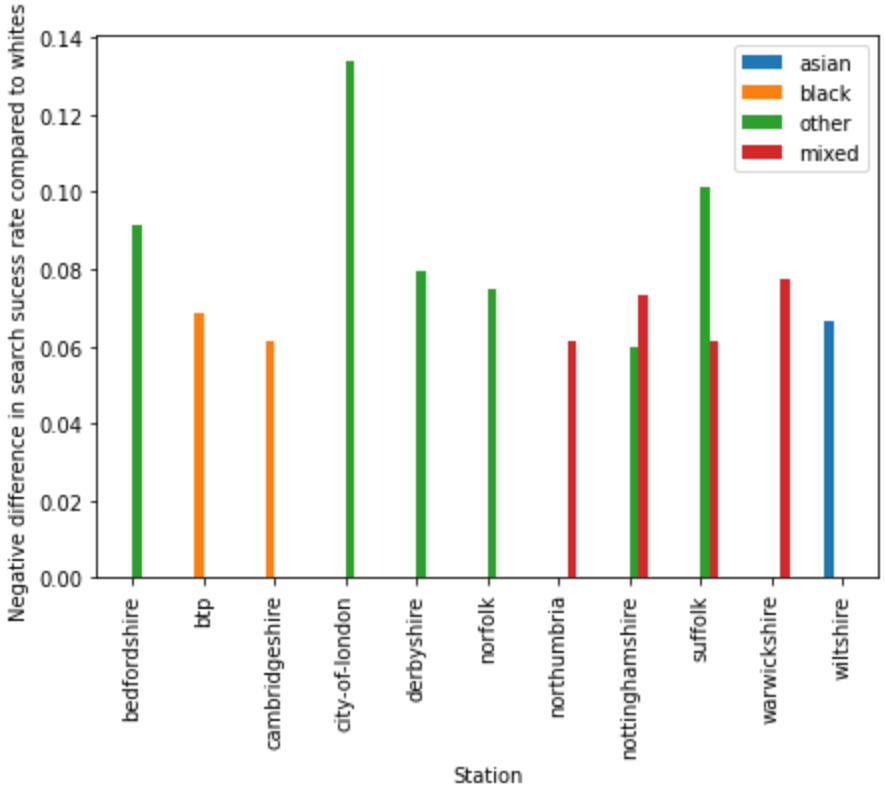
Stations for which difference in search success rate between non-white and white is greater than 5% (non-white are undersearched) :
'avon-and-somerset', 'cambridgeshire', 'cheshire', 'cleveland', 'derbyshire', 'dorset', 'gloucestershire', 'greater-manchester', 'hampshire', 'lincolnshire', 'north-yorkshire', 'northamptonshire', 'staffordshire', 'west-mercia'

Figure 2 : Positive difference (>5%) in search success rate between non-whites and whites for each station



Stations for which difference in search success rate between non white and white is lower than -5% (non-white are oversearched) :
‘'Bedfordshire', 'Btp', 'Cambridgeshire', 'City-of-London', 'Derbyshire', 'Norfolk', ‘Northumbria', 'Nottinghamshire', 'Suffolk', 'Warwickshire', 'Wiltshire'

Figure 3 : Negative difference (>5%) in search success rate between non-whites and whites for each station



Model technical analysis

Model parameters

The model was deployed using scikit-learn version 0.24.0, a package for Machine Learning with Python. The Random Forest Classifier was an instance of class “sklearn.ensemble.RandomForestClassifier”. The documentation of the classifier can be found in sklearn Documentation. We set the following parameters:

- min_samples_leaf=30 : minimum number of samples required to be at a leaf node
- class_weight="balanced" : Use the values of the target class to automatically adjust weights inversely proportional to class frequencies in the input data.
- random_state=42 : controls the randomness of the bootstrapping of samples and the sampling of the features.
- n_jobs=-1 : The number of jobs to run in parallel. (all processors)

The other parameters of the classifier were left with default values.

Before feeding the data to the classifier, we normalized the numerical data using the transformer “StandardScaler” from sklearn package. The categorical features were encoded with “OneHot”

class from “category-encoders” package version 2.2.2. To impute missing values the sklearn package SimpleInputer was used. The following transforming pipelines were created:

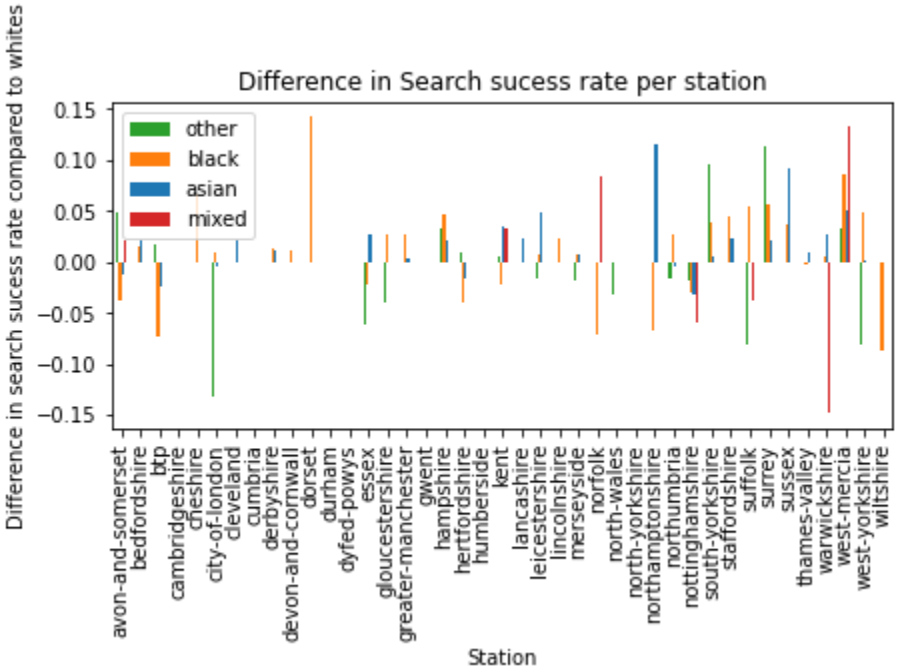
```
numeric_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='constant', fill_value=0)),
    ('scaler', StandardScaler())])

categorical_transformer = Pipeline(steps=[
('imputer',SimpleImputer(strategy='constant',fill_value='missing')),
('onehot',ce.one_hot.OneHotEncoder(use_cat_names=True,handle_unknown=
'indicator'))])
```

Discrimination metrics

Stations for which absolute difference in the precision between non-white and white is greather than 5% :

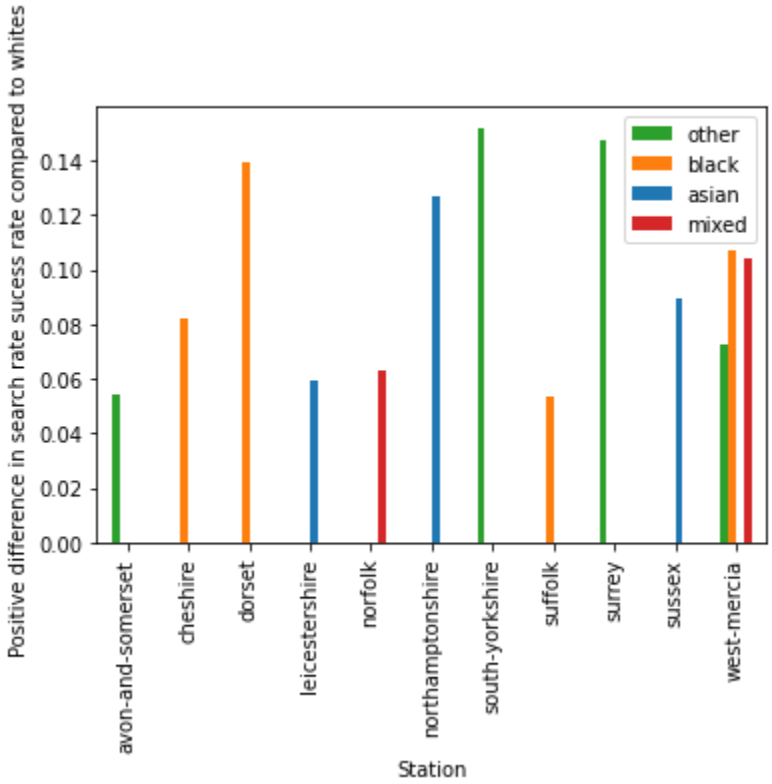
btp', 'cheshire', 'city-of-london', 'dorset', 'essex', 'norfolk', 'northamptonshire', 'nottinghamshire', 'south-yorkshire', 'suffolk', 'surrey', 'sussex', 'warwickshire', 'west-mercia', 'west-yorkshire', 'wiltshire'

Figure 1 : Difference in precision between non-whites and whites for each station



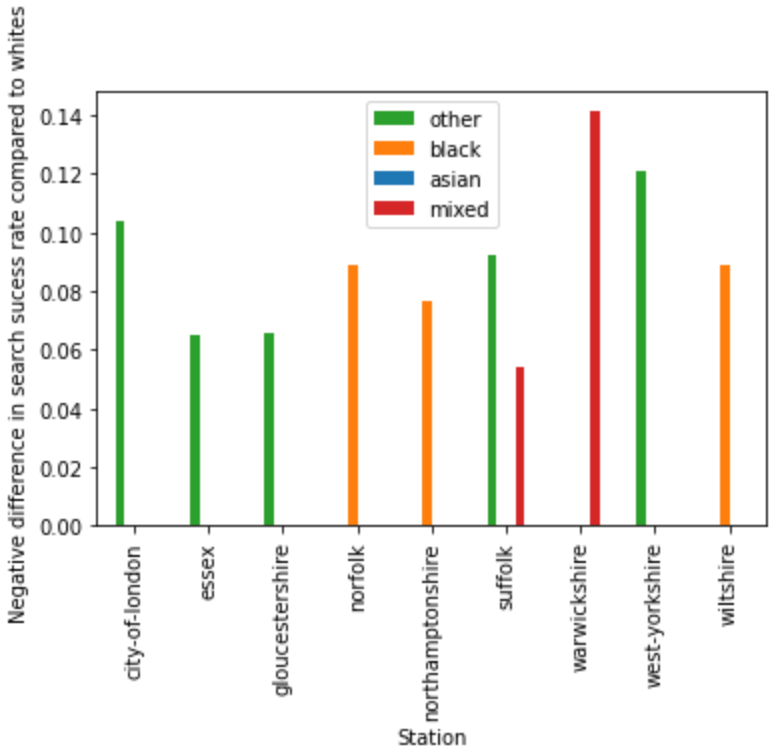
Stations for which difference in precision between non-white and white is greater than 5% (non-white are undersearched) :
'cheshire', 'dorset', 'norfolk', 'northamptonshire', 'south-yorkshire', 'suffolk', 'surrey', 'sussex', 'west-mercia'

Figure 2 : Positive difference (>5%) in precision between non-whites and whites for each station



Stations for which difference in precision between non white and white is lower than -5% (non-white are oversearched) :
'btp', 'city-of-london', 'essex', 'norfolk', 'northamptonshire', 'nottinghamshire', 'suffolk','warwickshire', 'west-yorkshire', 'wiltshire'

Figure 3 : Negative difference (>5%) in precision between non-whites and whites for each station


Negative difference in search sucess rate compared to whites
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
other
black
asian
mixed
city-of-london
essex
gloucestershire
norfolk
northamptonshire
suffolk
warwickshire
west-yorkshire
wiltshire
Station