# 

[**Client requirements**](#_1zetrph27kbs) **2**

[Summary](#_yhcjakxh17mu) 2

[Requirements clarifications](#_oq7yoe4kjuf1) 2

[**Dataset analysis**](#_d9jo2aheks9l) **3**

[General analysis](#_werogbweqbod) 3

[Business questions analysis](#_afmqws1f0aiu) 6

[Conclusions and Recommendations](#_bec7bbygurr7) 9

[**Modeling**](#_aa8l773uzfrm) **9**

[Model expected outcomes overview](#_qopili2ry7fk) 9

[Model specifications](#_9wqjxgqls5y5) 9

[Analysis of expected outcomes based on training set](#_wd5m1ybk1r7b) 10

[Alternatives considered](#_10qukq4k762o) 12

[Known issues and risks](#_ukq6fg56ovn0) 13

[**Model Deployment**](#_n5ss8dka3xzg) **13**

[Deployment specifications](#_93a9boam9d52) 13

[Known issues and risks](#_sai9k63u200y) 14

[**Annexes**](#_ir85ggj69yc) **15**

[Dataset technical analysis](#_1oh87ny9hvd6) 15

[Technical database summary](#_ldogfxk6gnii) 15

[Database visual overview](#_b0en15928vdq) 15

[Business questions technical support](#_je7d7xn4zwbh) 17

[Success search definition](#_ttmg2ba68ssd) 18

[Requirements verification in training database](#_49p6u9uk062i) 18

[Removal clothes analyzes by station](#_d608hsvxd7w) 19

[Model technical analysis](#_sizac7wphyik) 20

# 

# Client requirements

## Summary

The United Kingdom Department of Police is facing complaints about Stop and Search Operations and the aim with this report is:

* To analyze if there is evidence that any stations 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
* To develop and host an API endpoint for authorizing searches with the objective of:
  + Per station and per search objective, searches are performed only when there is more than 10% likelihood that the search will be successful
  + The success rate of the searches would not vary significantly between populations
  + Making sure no population is over-searched, which is defined as having equal success rates
  + API should be able to level the discovery rate without significantly diminishing the overall ability to detect offences

## Requirements clarifications

In order to translate the above expectations into measurable requirements, clarifications were made with the project’s sponsor and the below requirements were defined:

|  |  |
| --- | --- |
| **Business Requirement** | **Technical Requirement** |
| Minimum 10% success rate for searches per station and search objective | Precision >= 10% per station and search objective |
| Discrepancy between stations (average per station) would not be larger than 10 percentage points | Precision difference between stations would not be larger than 10 percentage points |
| No police station should have a discrepancy bigger than 5% for the search success rate between protected classes ethnicity and gender | Per police station, the precision difference between protected classes ethnicity and gender should not be bigger than 5% |
| No police station should have a discrepancy bigger than 5% for the Removal of more than just outer clothing rate between protected classes (age, ethnicity, gender) | Per police station, the Removal of more than just outer clothing rate between protected classes (age, ethnicity, gender) should not be bigger than 5% |
| API should be able to level the discovery rate without significantly diminishing the overall ability to detect offences | Maximum recall considering the requirements above |

*Table 1 - Business and Technical Requirements*

A search is considered successful if the search outcome is positive, and it is related to the search. For a technical definition, please see section Success search definition in the annexes.

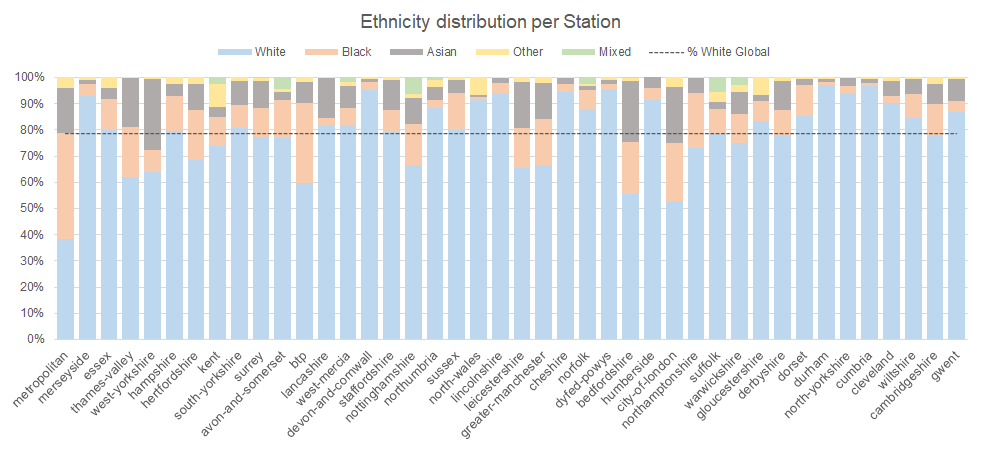
# Dataset analysis

## General analysis

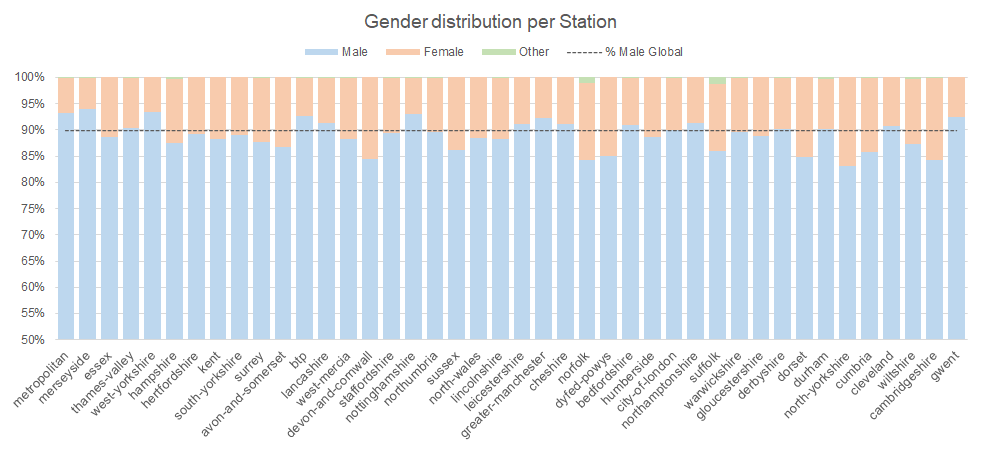
The database has around 660.000 searches made since 2017 by 42 different stations, where were collected:

* Person characteristics: age, gender, self-defined ethnicity and officer-defined ethnicity
* Search information: Date, latitude, longitude, legislation, object of search, indication if the search was part of a policing operation, type of search, station and indication of removal of more than just outer clothing
* Search result: Outcome of search and outcome linked to object of search, used to identify if a search was successful

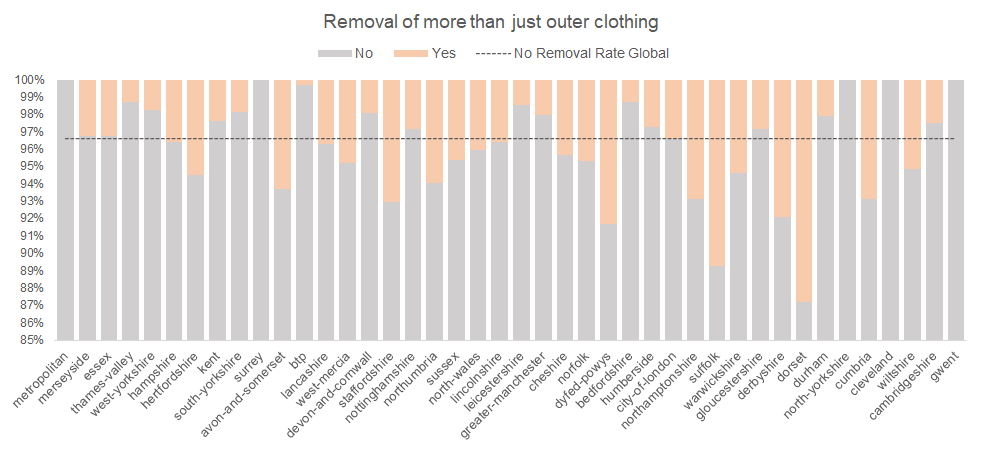
**Protected classes Overview**



*Graphic 1 - Ethnicity distribution per station*



*Graphic 2 - Gender distribution per station*

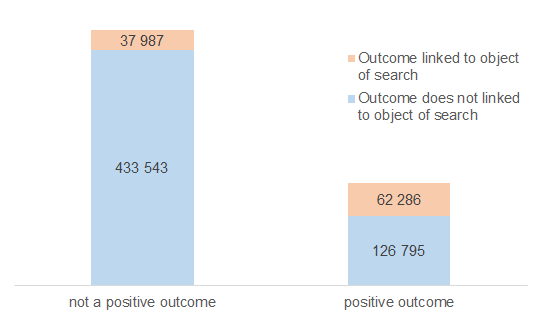


*Graphic 3 - Removal of more than just outer clothing distribution per station*

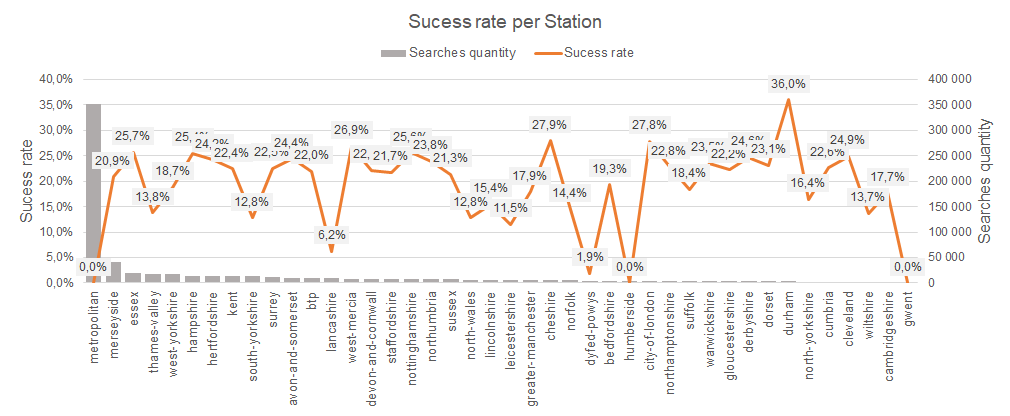
Please consult section “Technical database summary” in the annexes to see more exploratory data analysis.

**Successful search**

A search is classified as successful when it has a positive outcome, and the outcome is linked with the object searched. Thus, Graphic 4 shows that there are 62,286 successful searches.



*Graphic 4 - Positive outcome versus Outcome linked to the object of search*



*Graphic 5 - Success rate per Station*

**Data quality issues**

One important aspect of exploratory data analysis is the possibility of identifying data quality issues. When it occurs, it is necessary to be deeper analyzed to understand its root and also to be able to correct it and have coherent future data.

During this exploratory analysis the following problems were identified:

* Graphic 3 shows that for stations Metropolitan, Surrey, North-yorkshire and Gwent there is no information about removal clothing.

For the next analysis, these stations were not considered in the removal clothing analysis.

* Graphic 4 shows that there are 37,987 observations classified as linked to the object of search but without a positive outcome.

For the next analysis, these observations are being classified as not successful.

* Graphic 5 shows that the Stations Metropolitan, Humberside and Gwent have zero successful searches, while station Dyfed-powys has a very low rate.

For the next analysis, these 4 outliers stations are not being considered.

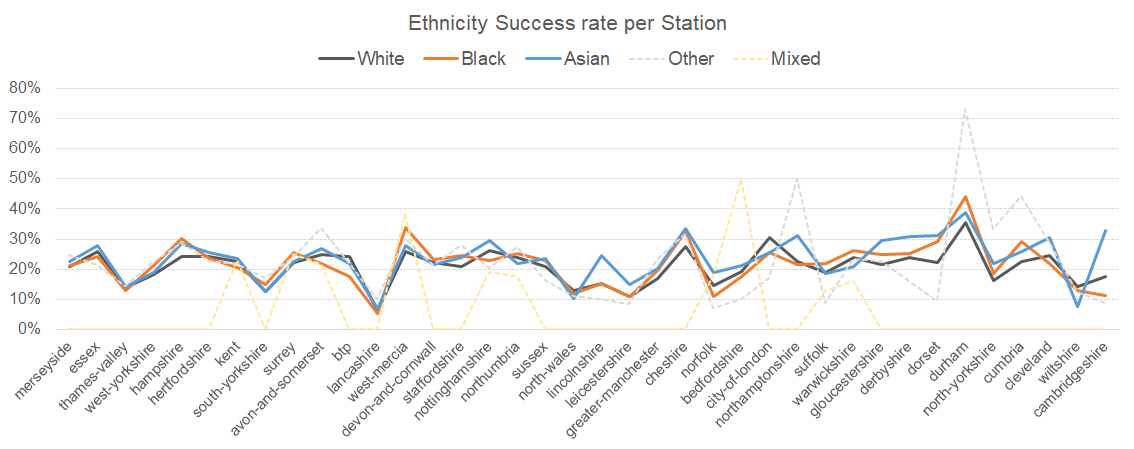
## Business questions analysis

The following analysis were made with the objective of looking for evidence that any stations may be discriminating on gender, ethnicity or age regarding who they choose to stop, and also who they ask for any clothing to be removed.

|  |  |  |
| --- | --- | --- |
|  | **% Searches** | **Success Rate\*** |
| **White** | 78.5% | 20.7% |
| **Black** | 10.7% | 20.9% |
| **Asian** | 8.2% | 20.3% |
| **Other** | 2.0% | 21.5% |
| **Mixed** | 0.6% | 20.9% |
| **Global** | **100.0%** | **20.7%** |

*Table 2 - Global success rate per Ethnicity*

For more details about success rate definition, please go to subsection “Successful search”.

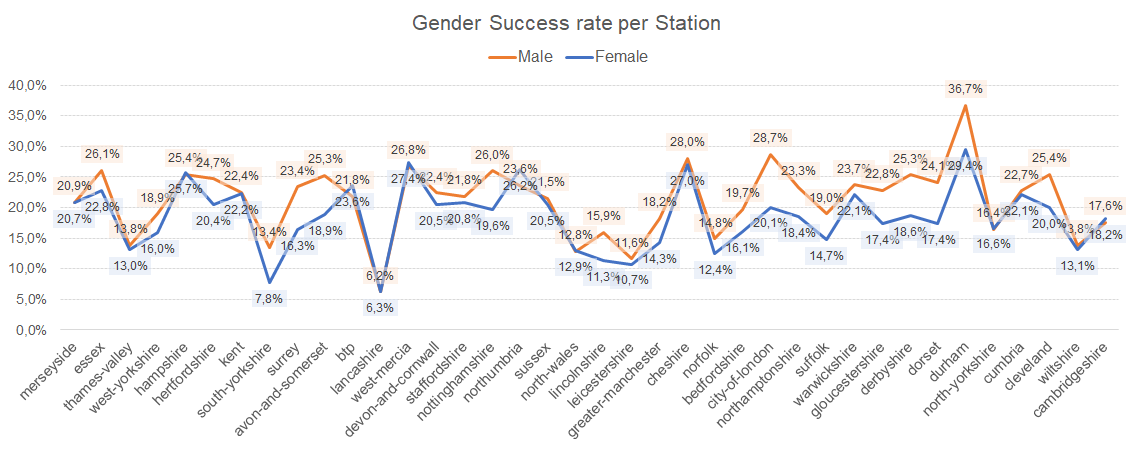


*Graphic 6 - Success rate per Station and Ethnicity. The stations are in descending order by number of searches.*

Analyzing Table 2, it is possible to consider that the discrepancy of success rate between ethnicities is low. The same behavior can be observed per station (Graphic 6). A major variability between ethnicities is observed only in the stations with less observation, which can be a result of small sample size.

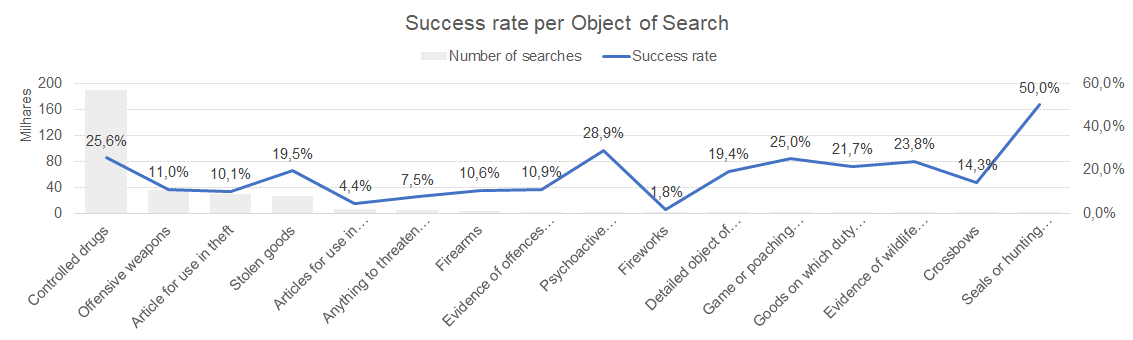
|  |  |  |
| --- | --- | --- |
|  | **% Searches** | **Success Rate** |
| **Male** | 89.8% | 20.9% |
| **Female** | 10.1% | 18.7% |
| **Other** | 0.1% | 17.6% |
| **Global** | **100.0%** | **20.7%** |

*Table 3 - Global success rate per Gender*



*Graphic 7 - Success rate per Station and Gender. The stations are in descending order by number of searches.*

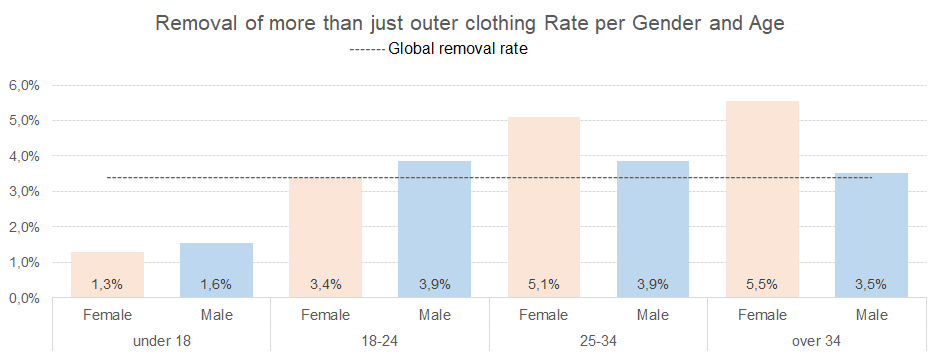
Analyzing Table 3, it is possible to see that there is a 2.2 percentage points difference between male and female success rates. This difference only increased in the stations with a small number of searches (Graphic 7).



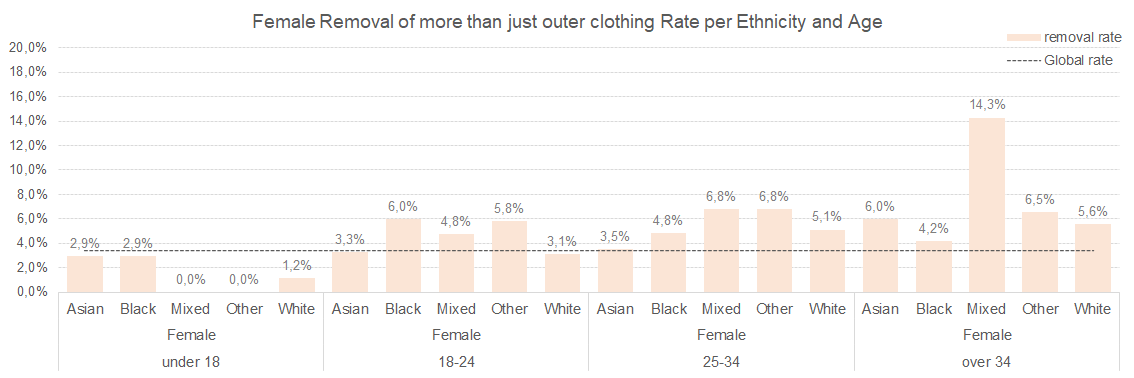
*Graphic 8 - Success rate per Object of search*

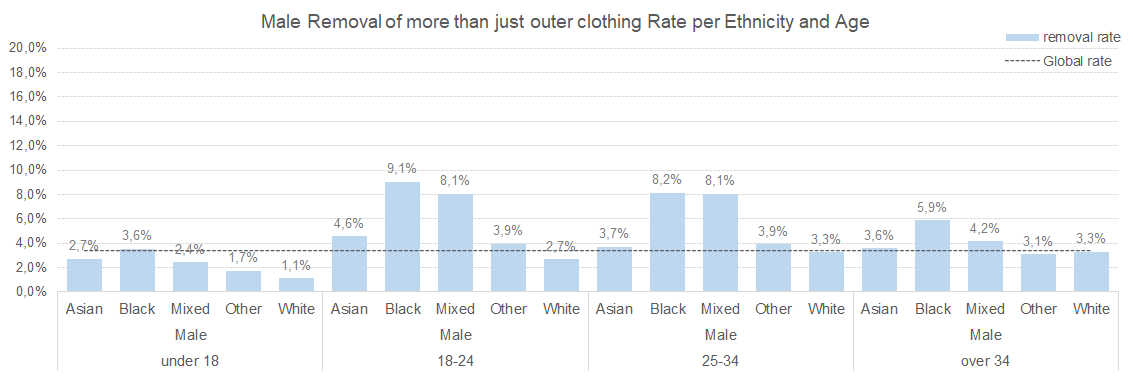
The success rate when the object of search is “offensive weapons” or “article for use in theft” (Graphic 8) is, approximately, half of the overall success rate (20.7%).

The global “removal of more than just outer clothing” rate is 3.4% and, by observing Graphic 9 and Graphic 10, it is possible to see that females over 25 years old are over asked to remove clothes in all ethnicities. Men from black and mixed ethnicities are also over-asked for clothes removal.



*Graphic 9 - Removal of more than just outer clothing Rate per Gender and Age.*





*Graphic 10 - Female and Male Removal of more than just outer clothing Rate per Ethnicity and Age*

## Conclusions and Recommendations

When the success rate is analyzed in a global perspective, it looks like that there is no reason for over searching on any specific ethnicity, as the success rates are quite similar (Table 2). The following stations have an over-search for some ethnicities that suggest the need of a deep analysis and special actions to correct it, as seen on Graphic 1: Metropolitan, Thames-Valley, Btp, Bedfordshire, City-of-London and West-Yorkshire.

Regarding whom are being asked for clothing to be removed, it is possible to see an overall over-asking for women with more than 25 years old and black and mixed men over 18 years old. Details by stations are available in the annexes “Business questions technical support”.

# Modeling

## Model expected outcomes overview

Several data science techniques, described in [Alternatives considered](#_10qukq4k762o) section, and good practices were applied in the model construction aiming to fulfill the requirements.

However, considering that the model learns from the dataset with imbalance subpopulations, and a clear bias in the removal of more than outer clothing, not fulfilling all requirements is a possible scenario. Especially regarding removal clothing that will continue to be a police officer decision, and also with objects of search where some objects already have a lower success rate. The good news is that new approaches can be studied, tested and implemented (out of this project scope) in order to reduce removal clothing ethnicities bias and improve the success rate for some search objectives.

With the dataset available it was possible to develop a model where the overall success rate increased 4 percentage points, passing from 20.7% to 25% with only 15% of diminishing of overall discovery rate. The success rate variation between subpopulations is higher than the expectation but still satisfactory.

For the proof-of-concept, if the subpopulations remain approximately stable, a similar result is expected.

## Model specifications

* Training set:
  + Observations from stations Metropolitan, Humberside, Gwent and Dyfed-powys were not used because of the atypical successful rate
  + The model was trained with a sample of 70% the train set and tested and evaluated with the other 30%
* Target: defined as detailed in the annexes “Success search definition”
* Features:
  + Type
  + Part of a policing operation - missing values were filled with False
  + Age range - “under 10” and “10-17” were combined as a category “under 18”
  + Latitude - missing values were filled with 0 because it is a value out of the dataset possible values
  + Longitude - missing values were filled with 50 because it is a value out of the dataset possible values
  + Legislation - was recategorized to:
    - Missing values were assigned to a “missing information” category
    - All categories different than “misuse of drugs act 1971 (section 23)”, “police and criminal evidence act 1984 (section 1)”, “criminal justice and public order act 1994 (section 60)” or “firearms act 1968 (section 47)” were reclassified to “others”
  + Hour - hour of the search (Date feature)
  + Month - month of the search (Date feature)
  + Day\_of\_week - day of the week (Sunday, Monday, …, Saturday) of the search (Date feature)
* Numerical features:
  + Latitude, Longitude, Hour and Month
* Categorical features:
  + The remaining features
  + All categories from each feature were transformed in a dummy feature (one hot encoding)
  + The model is prepared to accept unknown categories
* Estimator:
  + Random forest with maximum depth of the tree = 300 and minimum number of samples required to be at a leaf node = 1% of the train dataset. The remaining parameters were used as default according to scikit-learn 0.23.2
* Predicted class:
  + An observation is predicted as True when the predicted probability is equal or greater than 0.414.

## Analysis of expected outcomes based on training set

As was seen in previous analysis, the training set is imbalanced in the features that are desirable to level (like stations with 30 p.p. of different in success rate), some natural characteristics of the dataset are distant from the requirements (like some search objectives that have low success rate), and the challenge of not having a discriminatory request for removal of clothes, once it is a police officer decision and it will continue to be.

Although it is realistic to expect an overall increase in the success rate, some stations cannot fulfill the requirements, once the 5 p.p. difference is an ambitious rule. It is completely possible to obtain a difference greater than 5 p.p. but still acceptable in this complex context.

Table 4 shows a summary of the model performance. It was possible to level the success rate for most of the stations and protected classes. For those that didn’t fulfill, we can see that the average difference is acceptable, taking account how heterogeneous some stations are, and that the average difference remains between 0.9 p.p. and 3.9 p.p. higher than the expected (5 p.p.).

The average difference per Ethnicity for removing clothes is the highest, and it is hard to handle because the decision of asking for removal is made by the police officer. One possible alternative for this situation is, as a future step, to build a model to decide whenever a person will be asked to remove clothes. Removing this decision from the police’s hands is a possible way to avoid the possibility of biased decisions.

|  |  |  |
| --- | --- | --- |
| **Requirements** | **Subpopulation** | **Model Fulfilment** |
| Minimum 10% success rate for searches per station and search objective | Station | 37 stations fulfill the requirement |
| Search objective | 20 stations fulfill the requirement |
| Discrepancy between stations would not be larger than 10 percentage points | Station | 30 pp of difference |
| No police station should have a discrepancy bigger than 5% for the search success rate between protected classes ethnicity and gender | Ethnicity | 8 stations fulfill the requirement  Average difference: 8.9% |
| Gender | 20 stations fulfill the requirement  Average difference: 8.0% |
| No police station should have a discrepancy bigger than 5% for the Removal of more than just outer clothing rate between protected classes (age, ethnicity, gender) | Gender | 31 stations fulfill the requirement  Average difference: 5.9% |
| Age | 23 stations fulfill the requirement  Average difference: 8.4% |
| Ethnicity | 18 stations fulfill the requirement  Average difference: 15.4% |
| *Table 4.1 - Requirements fulfillment* | | |
| **Technical Measures** | |  |
| AUROC | 0.61 |  |
| Global precision | 0.25 |  |
| Global recall | 0.85 |  |
| Threshold to define True class | 0.41 |  |

*Table 4.2 - Technical Measures*

## Alternatives considered

Before the final version of the model, different estimators and features combinations were tested. The table below resumes some of these versions.

The final model was chosen based on the better combination of precision, recall and fulfillment of the requirements with the lowest average difference between stations.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Requirements** | **Subpopulation** | **Final Model** | **Alternative 1** | **Alternative 2** | **Alternative 3** | **Alternative 4** |
| Minimum 10% success rate for searches per station and search objective | Station | 37 stations fulfill | 37 stations fulfill | 38 stations fulfill | 37 stations fulfill | 37 stations fulfill |
| Search objective | 20 stations fulfill | 0 stations fulfill | 4 stations fulfill | 21 stations fulfill | 19 stations fulfill |
| Discrepancy between stations would not be larger than 10 percentage points | Station | 30 pp of difference | 35 pp of difference | 16 pp of difference | 35 pp of difference | 35 pp of difference |
| No police station should have a discrepancy bigger than 5% for the search success rate between protected classes ethnicity and gender | Ethnicity | 8 stations fulfill  Avg diff: 8.9% | 9 stations fulfill  Avg diff: 11.2% | 9 stations fulfill  Avg diff: 11.5% | 8 stations fulfill  Avg diff: 9.0% | 11 stations fulfill  Avg diff: 10.5% |
| Gender | 20 stations fulfill  Avg diff: 8.0% | 15 stations fulfill  Avg diff: 9.9% | 21 stations fulfill  Avg diff: 8.3% | 20 stations fulfill  Avg diff: 7.3% | 18 stations fulfill  Avg diff: 8.3% |
| No police station should have a discrepancy bigger than 5% for the Removal of more than just outer clothing rate between protected classes (age, ethnicity, gender) | Gender | 31 stations fulfill  Avg diff: 5.9% | 32 stations fulfill  Avg diff: 10.3% | 33 stations fulfill  Avg diff: 12.6% | 32 stations fulfill  Avg diff: 9.2% | 20 stations fulfill  Avg diff: 7.3% |
| Age | 23 stations  Avg diff: 8.4% | 23 stations  Avg diff: 9.1% | 30 stations  Avg diff: 8.7% | 30 stations  Avg diff: 8.4% | 23 stations  Avg diff: 7.7% |
| Ethnicity | 18 stations  Avg diff: 15.4% | 13 stations  Avg diff: 12.0% | 12 stations  Avg diff: 11.2% | 13 stations  Avg diff: 13.2% | 13 stations  Avg diff: 12.0% |
| AUROC | | 0.61 | 0.59 | 0.50 | 0.50 | 0.51 |
| Global precision | | 0.25 | 0.25 | 0.25 | 0.25 | 0.26 |
| Global recall | | 0.85 | 0.75 | 0.89 | 0.81 | 0.43 |
| Threshold to define True class | | 0.41 | default - 0.5 | 0.14 | 0.16 | 0.25 |
| Estimator | | Random Forest | Random Forest | LGBM Classifier | Gradient Boosting Classifier | Random Forest without sensitive features |

*Table 5 - Final Model and Alternative models*

## Known issues and risks

As the nature of police operations is dynamic and changes are frequently made, it is possible to observe subpopulations distribution and/or success rate changes in a small space of time. And significant changes like these can make the model not respond with the expected accuracy.

Another source of issue is usage of API. If the requests sent to API are not complete and valid as expected, the model will have less information and can predict an inaccurate response.

# Model Deployment

## Deployment specifications

The final model was developed and fitted on Python using pipelines, and its results were serialized into three files: pipeline.picke, columns.json and dtypes.pickle. In order to replicate the deployment, the following steps are necessary:

* Inside a python file:
  + Create a HTTP server using flask python library
  + The API needs to have two end points:
    - /should\_search/ that returns “true”/”false” if the person/vehicle should be searched
    - /search\_result/ receives the true class of an observation and update it into the database
  + For /should\_search/, are verified:
    - If the request has the observartion\_id key. If there isn’t, an error message is returned
    - If all the needed features are being sent: observation\_id, Type, Date, Part of a policing operation, Latitude, Longitude, Gender, Age range, Officer-defined ethnicity, Legislation, Object of search and Station. If some of them are missing or an unknown feature is sent, an error message is returned
    - Latitude and Longitude must be a number (can be integer or float) and it is allowed to receive missing for these features. If any of them is not a number, an error message is returned
    - Part of a policing operation must be a Boolean. If it is a string, an error message is returned
    - If the observation\_id already exists in the database, a message is returned informing it
  + For /search\_result/ it is verified if the observation\_id has a prediction to be updated with the true class. If there isn’t a prediction for the observation, an error message is returned
  + The model uses some transformed features that need to be done before the prediction of the new observation:
    - Using the feature Date, creation of the features hour (numeric), month (numeric) and day\_of\_week (Sunday, Monday, …, Saturday)
    - Apply lowercase to the categorial features Type, Age range, Legislation, Gender and Officer-defined ethnicity
    - Fill missing value of “Part of a policing operation” with False
    - Fill missing value of Legislation with “missing information” and recategorize Legislation into “others” when the value is different than “misuse of drugs act 1971 (section 23)”, “police and criminal evidence act 1984 (section 1)”, “criminal justice and public order act 1994 (section 60)”, “firearms act 1968 (section 47)”
    - Fill missing value of Latitude with 0
    - Fill missing value of Longitude with 50
    - Recategorize Age range “under 10” and “10-17” into “under 18”
  + To be possible to predict a new observation, it is necessary to deserialize the trained model using the files mentioned before
  + The database on Heroku is prepared to save: observation\_id, observation (full request that was received), outcome (Boolean field with the predict class), prediction (probability predicted) and true\_class (to save the true class sent by search\_result)

## Known issues and risks

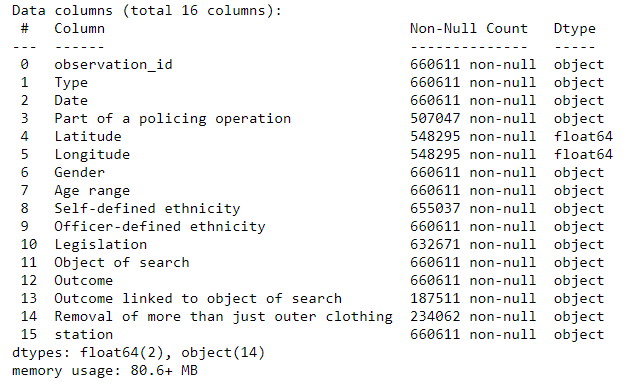
It was requested that it is necessary to accept new and unknown categories. Since almost all of the features are string and can receive a new category, it is not possible to check if the category that is being sent is valid (hard to define if it is a new category or a mistyping, for example). The API is prepared only to check the numerical and Boolean features (Latitude, Longitude and “Part of a policing operation”). This way, the API accepts any value for the categorical features, and it can affect the model accuracy in the course of time.

Once the API depends on the Heroku infrastructure, an eventual Heroky unavailability will reflect in an unavailability of the API.

# Annexes

## Dataset technical analysis

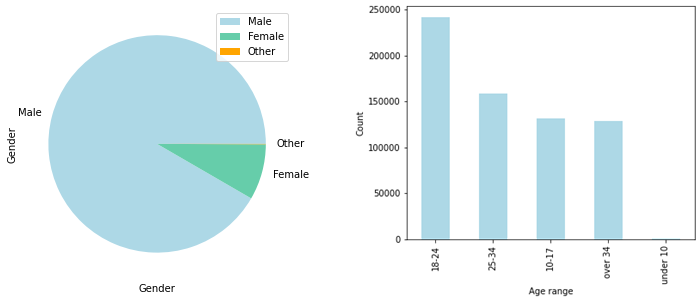
### Technical database summary



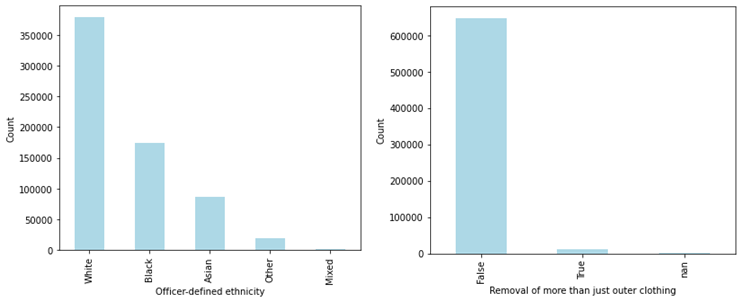
Some features have missing values that need to be handled in order to be possible to use the observation for modeling. According to Sponsor’s indications, the following adjustments were made:

* Part of a policing operation: the missing values can be interpreted as False (non-part of a policing operation)
* Outcome linked to object of search: the missing values can be interpreted as False (not linked to the object of search)
* Removal of more than just outer clothing: the missing values can be interpreted as False (no clothes removal) when the search type is “person search” or “person and vehicle search”

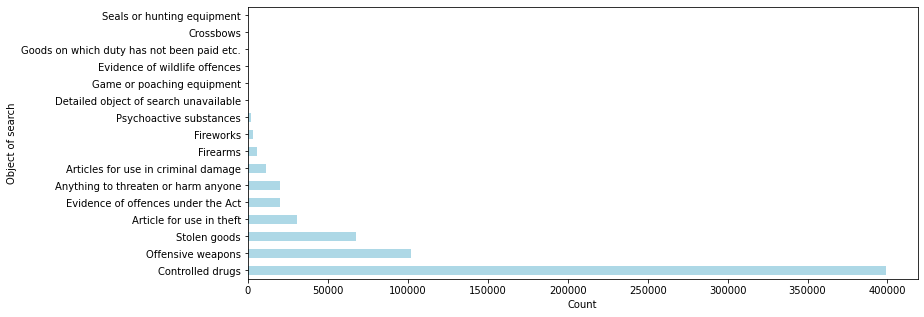
### Database visual overview



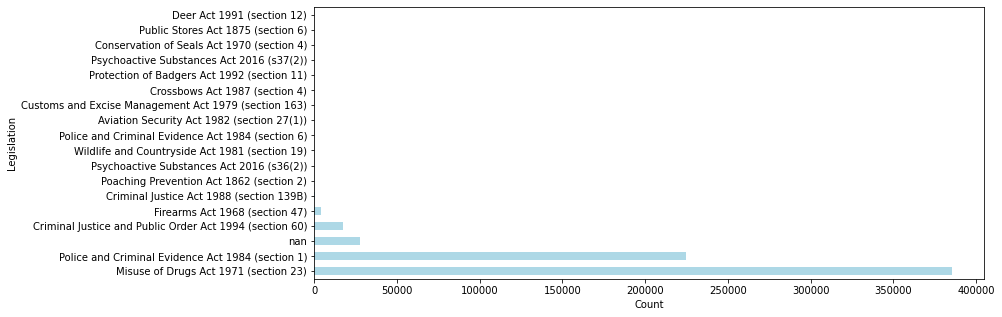
*Graphic 11 - Gender and Age distribution of the full training database*



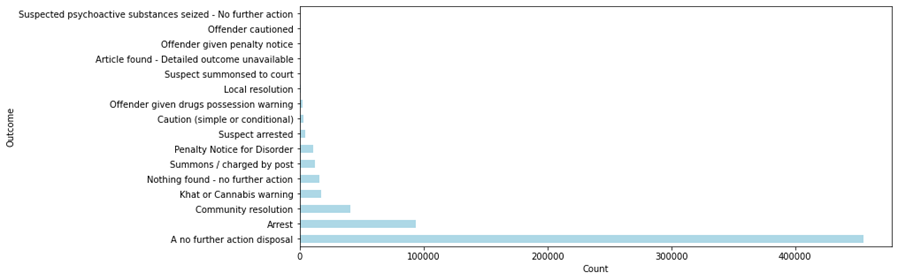
*Graphic 12 - Officer-defined ethnicity and Removal of more than just outer clothing distribution (full training database)*



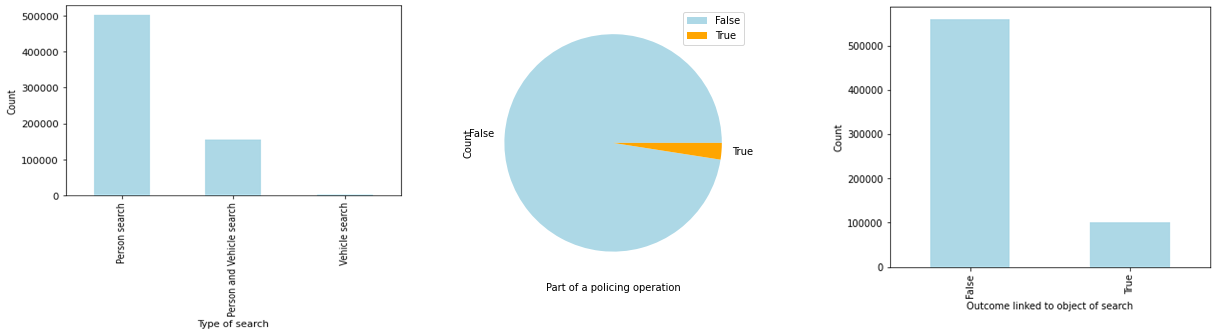
*Graphic 13 - Object of search distribution (full training database)*



*Graphic 14 - Legislation distribution (full training database)*



*Graphic 15 - Outcome (full training database)*



*Graphic 16 - Type of search, Part of policing operation and Outcome linked to object of search distribution (full training database)*

## Business questions technical support

### Success search definition

A search is considered successful if the outcome is positive and is related to the search. That is, the “Outcome linked to object of search” (missing values are considered as False) must be True and the Outcome has to be one of the categories:

* Local resolution
* Community resolution
* Offender given drugs possession warning
* Khat or Cannabis warning
* Caution (simple or conditional)
* Offender given penalty notice
* Arrest
* Penalty Notice for Disorder
* Suspected psychoactive substances seized - No further action
* Summons / charged by post
* Article found - Detailed outcome unavailable
* Offender cautioned
* Suspect arrested
* Suspect summonsed to court

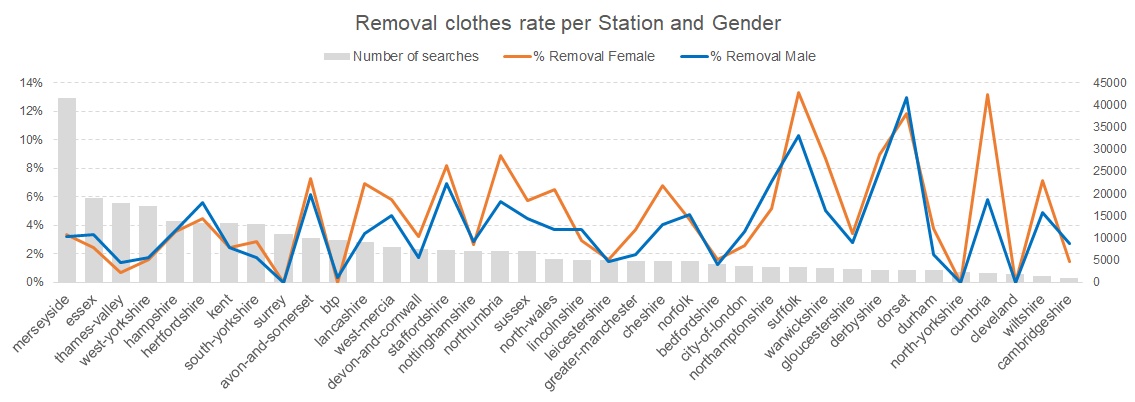
### Requirements verification in training database

To be able to set expectations about the model results, the requirements were verified in the training set.

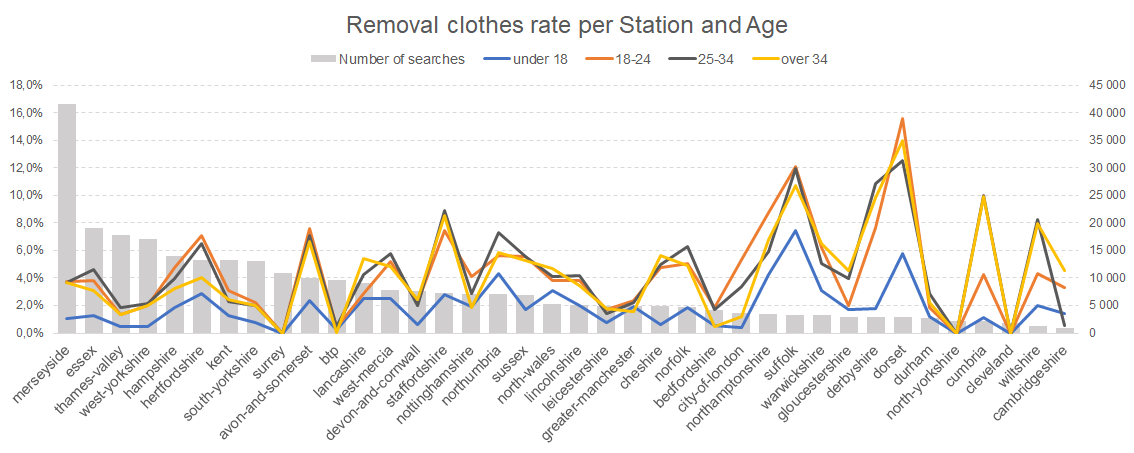
|  |  |  |
| --- | --- | --- |
| **Requirements** | **Sub population** | **Is this requirement currently fulfilled?** |
| Minimum 10% success rate for searches per station and search objective | Station | No - Lancashire has 6.3% success rate |
| Search objective | No - 3 objectives are lower than 10% and two of the top 4 in number of searches (“Offensive weapons” and “Article for use in theft”) have rate near by 10% |
| Discrepancy between stations would not be larger than 10 percentage points | Station | No - the biggest difference is 30 percentage points, and the standard deviation is 5.6% |
| No police station should have a discrepancy bigger than 5% for the search success rate between protected classes ethnicity and gender | Ethnicity | No - 18 of the 38 stations analyzed have more than 5 p.p. difference between ethnicity groups. |
| Gender | No - 10 of the 38 stations analyzed have more than 5 p.p. difference between genders |
| No police station should have a discrepancy bigger than 5% for the Removal of more than just outer clothing rate between protected classes (age, ethnicity, gender) | Gender | No - 1 station has 7 p.p. of difference |
| Age | No - 7 stations with more than 5 p.p. of difference |
| Ethnicity | No - 20 stations with more than 5 p.p. of difference |

*Table 6 - Requirements verification in training database*

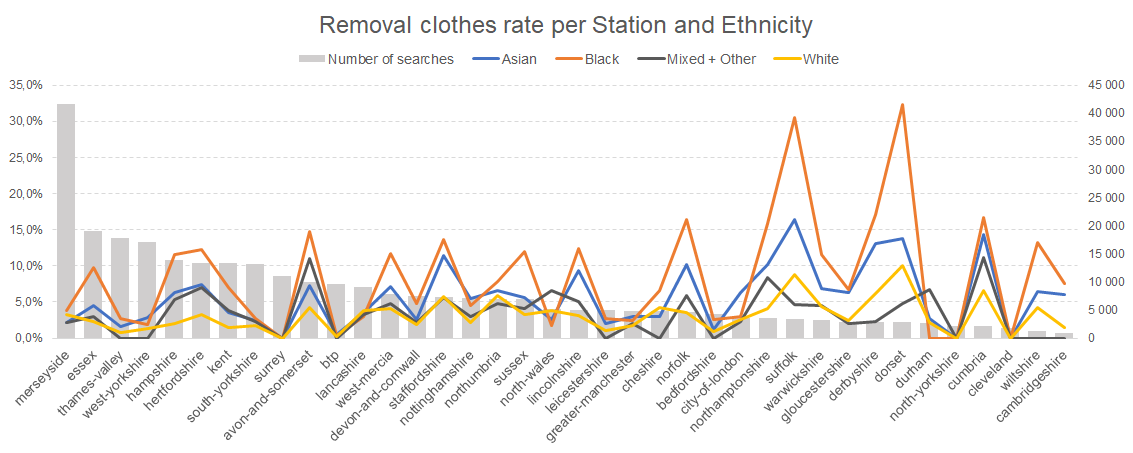
### Removal clothes analyzes by station



*Graphic 17 - Removal clothes rate per Station and Gender*



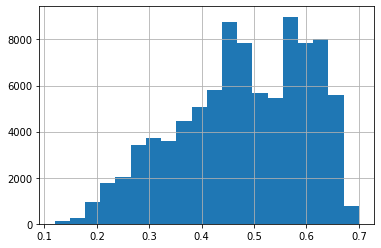
*Graphic 18 - Removal clothes rate per Station and Age*



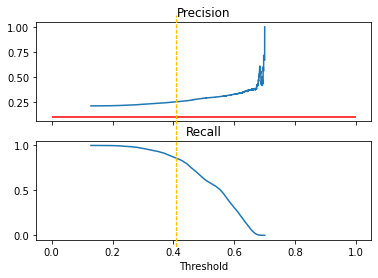
*Graphic 19 - Removal clothes rate per Station and Ethnicity*

## Model technical analysis

## As mentioned before, the final model was chosen based on a balance between good accuracy and requirements fulfilment with an acceptable average success rate difference between stations. Also, one more technical aspect was considered: the distribution of the probabilities of the model. The final model has a good distribution that allowed the usage of a threshold of 0.414 and avoid stations-objective search combinations with precision zero.



*Graphic 20 - Probability distribution of Final Model*



*Graphic 21 - Precision and Recall using different thresholds*

## The most important features for a successful search are some kind of legislation, the location of the search, the age of the person, and the hour and month of the search:

## 

*Graphic 22 - Feature importances*