TABLE OF CONTENTS

CLIENT REQUIREMENTS

Summary

Stop and Search operations are an important and controversial component of daily policing strategies in the United Kingdom (UK) recently receiving a fair amount of public criticism. Together with the help of the IT police Department, this project will address these concerns by investigating any evidence of wrongdoing, abuses or discrimination of minority or social sensitive groups by the authorities and will attempt to propose a decision-making model designed to aid police officers in selecting to search only when there is a high likelihood of success. In summary, the two major tasks this enterprise aims to deliver can be summarized in:

a) analysis of a dataset with records from a broad spectrum of police stations across the UK looking for potential evidence of discrimination towards minority groups, gender-related or age-related. Furthermore, we intend to address the highly concerning issue of police officers asking for the removal of more than outer clothing to certain groups, namely if these practices are more frequently applied to women of certain age brackets;

b) develop a model that can be easily translated into a 'search vs no search' advice service regarding potential stops and searches of individuals. The proposed model should be comprehensive and general enough to be applicable to all stations nationwide without significant differences in discovery rates between minority groups.

Requirements clarifications

The main goal of this project is to yield a model capable of uniformizing the decision to perform searches whilst trying to minimize any bias towards the different ethnic, gender and age groups within a certain police station.

Although many possible types of offences might be incurred in this project we will frame the task as a binary classification problem where observations in the training data are considered successful if the outcome is positive[1], i.e, observations where the model should predict infractions of law are likely to be occurring, and this outcome is related to the search.

We are asked to reduce the number of cases where a search does not result in a successful outcome, which, in other words, means that for the same number of people stopped the model should give a higher frequency of positive results. A general measure like accuracy is not suitable for this problem as we are dealing with an imbalanced dataset but this requirement can be translated into an analytic measure such as precision in increasing the discovery rate. However, by maximizing this it is only natural that the ability to dismiss potential offences is increased as well. This should be measured by the sensitivity/recall of the constructed model and a reasonable balance between them must be proposed. These metrics will be used to identify how many stations comply with a tolerance level of discrepancies within each group and each police station and achieve a balance between a unified policy and local possible differences. Ideally, we would seek to get the majority of stations with no more than 5-10% difference of successful discoveries within groups and between stations.

DATASET ANALYSIS

General analysis

For the task at hand, the IT Department has made available a dataset comprising 660 661 stop and search events spread throughout the country. This dataset is composed of features that can be used for modelling - 'Type', 'Date', 'Part of a policing operation', 'Latitude', 'Longitude', 'Gender', 'Age range', 'Self-defined ethnicity', 'Officer-defined ethnicity', 'Legislation', 'Object of search', 'station' - by features which will be used to build a classification target given there isn't a specific 'target' or ' label' feature - 'Outcome', 'Outcome linked to object of search' - and by the feature 'Removal of more than just outer clothing', to be used in the analysis only. As previously mentioned a search is considered successful if the outcome is positive and is related to the search. Except for 'Latitude' and 'Longitude', all the others are categorical variables - you can find a full breakdown of features and entries numbers in Annex. 1.

Data entry issues

There seems to be some issues regarding data entries although in general most of the dataset is quite well populated. Maybe not surprisingly, the London metropolitan station accounts for little more than 50% of all the data points in this dataset.

In regards to missing data across all stations, the columns 'Outcome linked to object of search' and 'Removal of more than just outer clothing' are the most affected (Annex. 2). A more granular analysis by police station allows for in depth appraisal of the reality of record keeping per station (Annex 3). Some stations are top exemplar - essex, suffolk, sussex, northamptonshire, norfolk and gloucester - with near 0% of data missing whereas other stations like thames-valley, lancashire and dyfed-powys have severe data keeping records with more than 30% of data missing, particularly in respect to geolocation, information if it was part of police operation and the two above mentioned categories. The most severe issue pertaining to missing data is related to the feature 'Outcome linked to object of search' in the metropolitan station. There is no data at all, which will disallow these data points to be included in any modelling attempt and impair any statistical analysis related to outcome in this important area of the country. The station gwent and humberside also have no data in this feature.

Another worrying fact related to this project is the almost entire lack of report regarding the removal of outer clothing. Some stations like gwent, cleveland, north-yorkshire, metropolitan and surrey do not report this information at all and many others have less than 50% reporting in this category.

For some columns it makes sense to infer the missing data, for example in 'Outcome linked to object of search' any missing values are most likely to be False since officers tend to forget to go back to the application.

There seems to be a substantial amount of records where 'Outcome' is negative but the column 'Outcome linked to the object of search is True. This indicates potential compromise when training any model as these data points may be of no use. In figure 3 we can clearly see that some stations (warwickshire,btp,west-yorkshire and derbyshire) have more than 50% of observations with this mismatch.

As mentioned the features ‘Outcome’ and 'Outcome linked to object of search' will be used to build our target labels for binary classification. The latter is a boolean feature with True or False values. The Outcome feature is a wide range of consequences from arrestas to penalties or mere caution but the vast majority is that no action was taken. After constructing the target feature we achieve an imbalanced dataset where the positive outcome accounts for approximately 20% of the observations. By removing the stations where the target is not possible to be constructed (metropolitan, gwent and humberside) the final data set is only 46% of the original data.

Gender, Ethnicity and Age groups distribution of stops can be seen in Annex 4. The Gender feature has 3 categories: Female, Male and Other. Given that the representativity of ‘other’ is extremely low in the dataset (less than 0.001%) we will exclude this from our analysis and modelling. There is an over representativity of Male compared to Female. Females account for little more than 8% but if the metropolitan station is removed then it increases slightly to more than 10%. In regards to Ethnic groups we have 2 different features: Officer defined and self-defined. For modelling purposes it only makes sense to use the Officer-defined feature as this is the only information present prior to any decision making. By far the majority is composed of ‘white’ subgroups making white man the most stopped.

The age brackets recorded show a majority of young adults between 18 and 34 years old. Possibly of concern is the fact that there are 384 occurrences of children aged less than 10 years old.

In regards to ethnicity we also investigated if there was any significant discrepancy between Officer-defined and self-defined reported values. Although not significant, it seems that officers tend to dismiss the categories 'Mixed' and 'Other' , distributing them more across 'White' and 'Black' categories (Annex 5).

According to the data, only 2.5% of the observations were conducted in the context of a pre-planned police operation whereas the overwhelming majority is mainly ad-hoc and unexpected. Also, most of the searches are directed to people only (76%) and 24% are person and vehicle searches. Searches in vehicles only are negligible.

The values in columns ‘Legislation’ and ‘Object of search’ are seemingly correlated and might actually be redundant. An overview of the types of values one might encounter here can be seen on Annex 6. Overwhelmingly the reason why people are stopped is for search of possession of drugs or drug related issues.

The Latitude and Longitude features are present in most of the data entries. An analysis of the data points conclude that the observations are spread across the country including Wales, Scotland and Northern Ireland and not only England. Although not checked, we assume that there is a direct relation of these geolocation data with the nearest police station.

The data on ‘Removal of clothing’ is very sparse with more than 420 thousand observations lacking in the field. Nevertheless it seems to occur in 4.5% of stops and searches carried out.

The Date feature seems to be in string format which can then be transformed into a datetime object in order to perform feature engineering whilst modelling.

Business Questions analysis

The main question this report seeks to investigate is potential discrimination towards age, gender or ethnic groups across the stations and nationwide. To do this we looked at the current distribution of discovery rate of offences in these subgroups and across stations (excluding the ones where the target label couldn't be attributed). The global discovery rate is at 20% as discussed but it varies greatly across the country (Annex 7).

Gender - In general, although men are stopped a lot more than women the discovery rate is usually around the same for both groups, i.e, from the stopped people, the same proportion of men and women have been caught committing offences. Nationwide, the percentage of men and women stopped who are actually found to be practising some offence is around 20% (Annex 8). Some stations have higher discovery rates for men like city-of-london, Durham or surrey where the difference for women is around 7-9%. Other stations report a slightly higher number of discoveries for women like Northumbria or the transport police but the difference is only about 2% compared to men.

Ethnicity - According to the UK census data from 2011[1] the frequency of white people in the UK is around 87%, black people around 3% and asians 5%. This is even more evident in Whales, Scotland and Northern Ireland where the representativity of white people surpasses the 95% mark. If we consider all the data it is very clear that black and Asian minorities are overrepresented when compared to census data - 26% for black and 13% for asian. However, when we exclude the metropolitan station get a lot closer to census with 10% for black and 8% for asians only. When analysing the discovery rate, excluding the metropolitan station, we see that the proportion of offences is very similar in all ethnic groups. These statistics can be found on Annex 9 where we went more granular investigating if there was any discrimination not only in offences but also arrests, penalties or simple caution issuances as these have very distinct levels of gravity. We couldn't find any obvious evidence in regards to what constitutes our positive class. Within stations we find some the maximum discovery rate discrepancy between ethnic groups to be 22% but the average difference nationwide is 7% (see annex 10)

Age range - people in the age range are stopped at a higher rate and also seem to have a higher discovery rate than the other age groups. However the difference between these 2 metrics perhaps implies that there might be a discrimination issue with this age range.We also noticed that people over 34 have been issued penalties a lot more frequently than the other groups that are more likely to get issued mere caution notices (annex 11). When looking at the distribution of offences per station it seems that there aren't significant differences on all stations.

Removal of more than outer clothes - It is very infrequent for removal of outer clothing to be done. It is slightly more likely to happen to women than men - 4.1% vs 3,4%. It was interesting to verify that to the people asked to remove more than outer clothing, the discovery rate actually increases to 30%. Addressing the specific question of this being asked to certain

[1] https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/2011censuskeystatisticsandquickstatisticsforlocalauthoritiesintheunitedkingdom#part1/r21ukrttableks201ukladv1_tcm77-330436.xls

females' age groups rather than others, we verify that indeed it is a lot more frequent for women in the range over 34 to be asked to remove outer clothing (annex 12).

Conclusions and Recommendations

One of the main take-aways from analysing the data is the lack of data. The data entry issues have to be addressed particularly in the feature ‘Outcome linked to object of search’ in the metropolitan station. There are 9 stations where more than 20% data entries are missing and some where some features are totally absent. The feature parto of a policing operation seems to be the one where most stations fail to register data. We believe this could easily be corrected. Also of concern is the lack of records for the removal of more than outer clothing nationwide. Another issue is to clarify what the value ‘other’ means in the gender feature. Is this just a data entry issue or, in fact, the individual identified as ‘other’ and explore if this feature should be expanded to include more denominations of gender as many people do not see themselves as either female or male.

A distinct point when looking at the metropolitan station has a very worrying distinction, as 40% of stop and search is done on the black group and 17% on asian which is way above the national average for those groups. This should raise concern and grant future study.

Also, it seems worrying that a number of stops and searches were conducted to under 10 years old children particularly in London and Manchester.

The discovery rate is very distinct from station to station, hopefully a unified model for bridging this gap can be achieved. On a per station basis there are some where differences in groups are also very wide and we seek to close that gap as well.

Interestingly it seems that people over the age of 34 are more easily penalized with penalty fees instead of cautionary notes when compared with the other age groups, this should also be pondered and considered for officers to be more vigilant when issuing fees.

MODELING

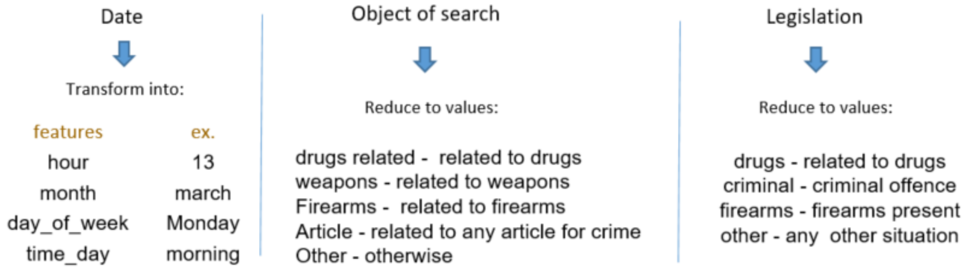
Model expected outcomes overview

The model developed and improved over time will attempt to minimize the discrepancies in discovery rate across the ethnicity, age and gender subgroups. In the data that was made available we couldn't identify any major differences in said groups albeit in some particular stations. But evidently we also do not know which individuals were left unsearched that potentially were breaking the law in some way. With this data it has been very challenging to unify all stations in a single policy model whilst keeping the identified potentially problematic groups with no discrimination. We would thrive to achieve an overall discovery rate higher than the original data (20%) which would translate into an approximate global 28% success rate. It has been challenging to unify the rates within groups. At the moment our model will yield some significant differences in discovery rate across different stations. The global discovery rate of people that are stopped is expected to be around 26%. We believe that an acceptable difference rate within stations should not exceed 10% within any subgroup but this is proving very difficult to achieve perhaps because the available features are not that representative of particular culprit idiosyncrasies of offenders. In some cases we think that we actually have not improved upon the existing stopping criteria, for example in regards to the gender sub group, but maybe we should, in the future, look to refine the models per station. We tested the best threshold for search and concluded that a good balance is that a person will only be searched if the model yields 50% likelihood of success. We will be monitoring variables such as maximum and average differences in subgroups nationwide and will attempt to level these metrics with future refinements to the model. In regards to people that should be stopped, i.e., of the people that are actually 'offending' it is expectable that we get a high rate of stops around the 0.74 mark. This metric should also vary significantly across the stations. The maximum differences expected regarding the targeted sub groups as well as other potential metrics can be found in annex 10.

Model specifications

The proposed model will make use of some features as they are, discard some and use others to perform feature engineering by creating new features. They are all categorical features. The non-categorical features are Latitude and Longitude which correspond to geolocation coordinates. These are not used in the model.

Features used: 'Type', 'Date', 'Part of a policing operation', 'Gender', 'Age range', 'station', ‘Legislation’, ‘Object of search’
Features not used: ‘Latitude’, ‘Longitude’,'Self-defined ethnicity',
Feature Engineering:



Date

Transform into:

features	ex.
hour	13
month	march
day_of_week	Monday
time_day	morning

Object of search

Reduce to values:

drugs related - related to drugs
weapons - related to weapons
Firearms - related to firearms
Article - related to any article for crime
Other - otherwise

Legislation

Reduce to values:

drugs - related to drugs
criminal - criminal offence
firearms - firearms present
other - any other situation

- Consequently, eliminate original features ‘Date’, ‘Legislation’ and Object of search’
- If ‘fitting’ the model with new training data exclude the samples with value ‘Other’ at feature ‘Gender’

The prediction of the model is a binary output of true or false for the stop and search to be conducted. To ‘fit’ or train a replication of this model the target feature has to be created. The ‘True’ label occurs when both ‘ Outcome’ and ‘Outcome linked to object of search’ are positive. For which values of ‘Outcome’ are considered positive please see Model technical analysis in annexes.

The model must be able to deal with missing values and categories that were not previously seen by the model on the training set.
Coding of all features categorical values using a one hot encoding strategy.

The algorithm used is a binary classification, non-linear model Random Forest with the following general specifications (for more details check Model technical specifications):

maximum depth of branches	15
number of trees	100
maximum leaf nodes	10
Class weight	balanced

Probability threshold used for predictions is above or equal to 0.5.

Analysis of expected outcomes based on training set

A balance between global discovery rates and global recall has been difficult. With a tolerance of 10% difference between groups we hope to achieve a global precision of 26% with the current model and a specificity or overhaul recall of about 75%.

Within the above mentioned tolerance we expect the current model to have ¾ of stations being compliant in regards to ethnicity with a maximum difference being 29% but the average below 10%. This difference was observed in the station cambridgeshire where the number of observations is very small accounting for only 876 samples of the total set. The gender subgroup should also perform relatively well with 32 police stations within the tolerance for discovery rate and a maximum of 18%. The sample size per police station does not seem to have had any impact when considering this group. There is a relatively narrow range of difference 10.5 to 18% when it comes to gender and it does not seem to correlate with the specific sample sizes.

The biggest problem will be the Age group where 32 stations do not comply with the tolerance. Most likely because indeed the 18-34 group is overwhelmingly searched and has a higher discovery rate and the under-10 group is under-represented. However the global average difference is only 14%. We will be monitoring this metric very closely and please refer to annex 10 for details of monitoring variables.

Based on this training set and experiments with randomized search for hyperparameter tuning we believe it will be very difficult to fulfill all the requirements namely to reduce sub groups potential for bias and achieve a nationwide effective protocol. Maybe a more targeted local approach instead of a unified policy would yield better results but at least the officers would be guided by an auditable model that can be improved over time.

Alternatives considered

Several iterations were tested with different ideas for feature engineering as well as different approaches to model selection and hyperparameter tuning. The optimization process was very focused on a balance between precision and recall both on the type of algorithm as well as on the corresponding parameters. The metric of choice for optimization was AUROC done with a strategy of randomized search. Only then one would test these parameters upon evaluation of our chosen business requirements metrics considering the balance tolerance between false positives and false negatives. We did investigate different decision making thresholds but in the end settled for 0.5 probability.

In regards to the algorithms studied, we tested linear (logistic regression) and non linear algorithms (Random Forest and Decision Trees) and decided on the Random Forest option although the differences were very subtle. On the feature engineering front we decided not to use the geo-location data because it was most likely related to the station's location so it would be redundant. We tested without using any feature building from the 'Date' feature but the results were not promising. We also attempted cost-sensitive learning on scikit learn library as it is available and had much better results by opting for a balanced class_weight. We did not investigate over sampling of the target data but did try random under sampling of the majority target class however the results were not better in reducing overhaul discrimination in number of police stations. Also, the stratification of sub groups was attempted but seemed to impact substantially the occurrence of discrimination on other groups, again in terms of number police stations not complying with the 10% tolerance in difference within groups.

Known issues and risks

We do not expect that the overall precision and recall nationwide to be far from our results in the out-of-sample tests. However we do see the potential for specific stations to broaden their already established biases. For example,we opted to remove the value 'other' in Gender from the training dataset so this could have an impact in the Gender subgroup. On our initial analysis, despite the massive difference in men being stopped, the discovery rate was pretty much the same for both groups. With this model this might change in some stations.

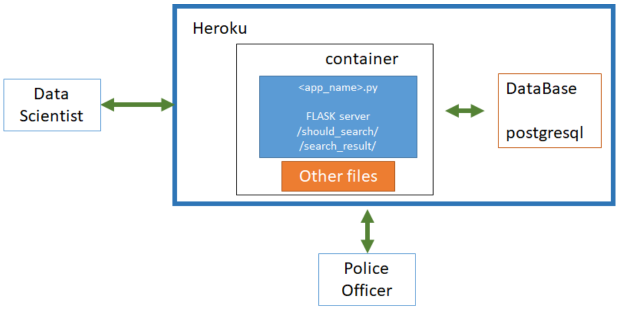
The amount of features made available do not seem to be giving any particular signals that would be useful to improve on meeting the requirements. Also the absence of the metropolitan station from the dataset given its data entry issues will probably have a tremendous impact on our modelling as 50% of data in this dataset was not used for modelling. Also, the local demographic, economic and social data specific to different locations is absent from these data and will, most likely, have an impact on our results. It is my understanding that if a unified nationwide approach is required than more granular data both from the individual as well as from the location characteristics is desirable.

In addition, it is important to note that the majority of reasons and the majority of issues identified in these stops and searches are related to a handful of factors of which the biggest problem being drugs issues. If, for some reason, the paradigm or distribution of offences changes, the model might not perform well. For example in the context of the current pandemic lockdowns and restrictions people might be prohibited from even walking on the street or gathering in large groups or may be stopped and fined for not wearing a mask. The current model does account for unknown, unseen potential offences so one would not expect a total lack of performance but these types of unforeseen unlawful events would probably impair the model's ability to yield good results if the change in the features becomes permanent for the long term. This would obviously require revisiting the modelling process and adaptation to the new paradigm.

MODEL DEPLOYMENT

Deployment specifications

Our application is built in python 3 and associated libraries using a PostgreSQL database resting in a cloud platform named Heroku. The main architecture can be translated into the following:



The app code is deployed inside a docker container that should be set to install all the required python libraries upon deployment. The application contains a file named <app>.py within which a server was built using a micro web framework named Flask. This server (API) is composed of two ['POST'] endpoints:

/should_search/ - returns the prediction of the model

/search_result/ - updates the database with the correct outcome

It is within <app>.py that we code the computational logic for connecting with the database, deserializes the model pipeline and makes use of any extra package such as any custom made transformer.

This code should also account for verification of valid inputs in any request and reject the request if any invalid parameters, missing values occur or if the observation_id has already been registered. However, it should accept 'null' values.

The application connects to a database (PostgreSQL) which is a service offered inside the Heroku platform.

An easy way to deploy the model would be to create a file with all the necessary components of the API and connect the Heroku platform to this folder through git. The main components inside this folder should be:

. <app_name>.py : code with models and endpoints

. Dockerfile: to set a container automatically installing all python necessary dependencies and libraries

. Requirements.txt: list of all necessary python libraries to run the app and model

. Heroku.yml: reference Dockerfiles in a heroku.yml file, set the app's stack to container and easily created and install the docker image

. Any serialized files in pickle or JSON format that the code requires

Known issues and risks

In terms of issues related to requests to be executed to the API, we believe that the system will consider invalid any request with missing values, missing features. Nevertheless, it will be able to accept null values and deal with new values for current features. The way it incorporates new 'unseen' values in the prediction computation is by giving the value 0 to all the dummy features that were created for this observation. This is far from ideal but it will avoid the system from crashing.

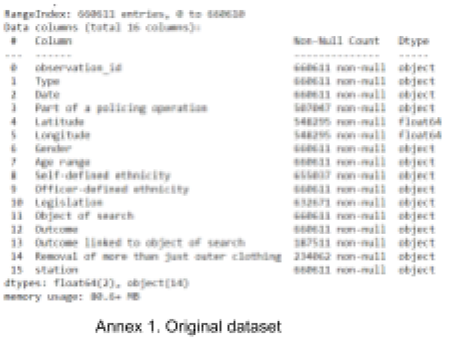
However, we do not know how the observation_ids are currently attributed so it is possible that, if an observation_id is repeated from, for example, some other station, that the observation_id will not be accepted for prediction since it is already recorded in the database.

The system will also not be able to run if extra (new) features are present in the request. In regards to the time of execution, given that the model is not very complicated, we wouldn't expect the prediction to take long to be computed and retrieved.

In addition, the smooth running of the system depends upon the server and databases in which it is implemented. Situations like the server being offline do happen and there is no backup for this at this time. Similarly, the database which also runs on the heroku server could suffer from the same problems. In reality, the database is an even bigger problem as it could potentially suffer a catastrophic deletion of our observations. Our code for the application is easily transferred to another server but there is no database backup from our end. It should be considered the creation of frequent backups of the database.

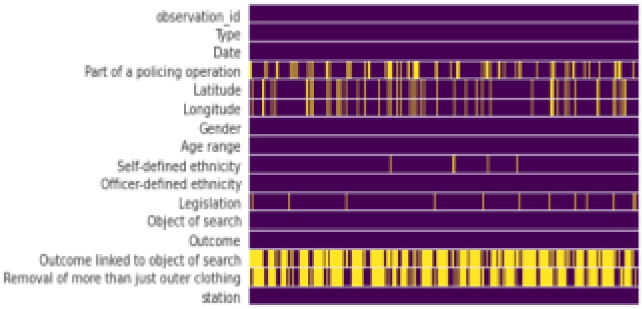
ANNEXES

Dataset technical analysis



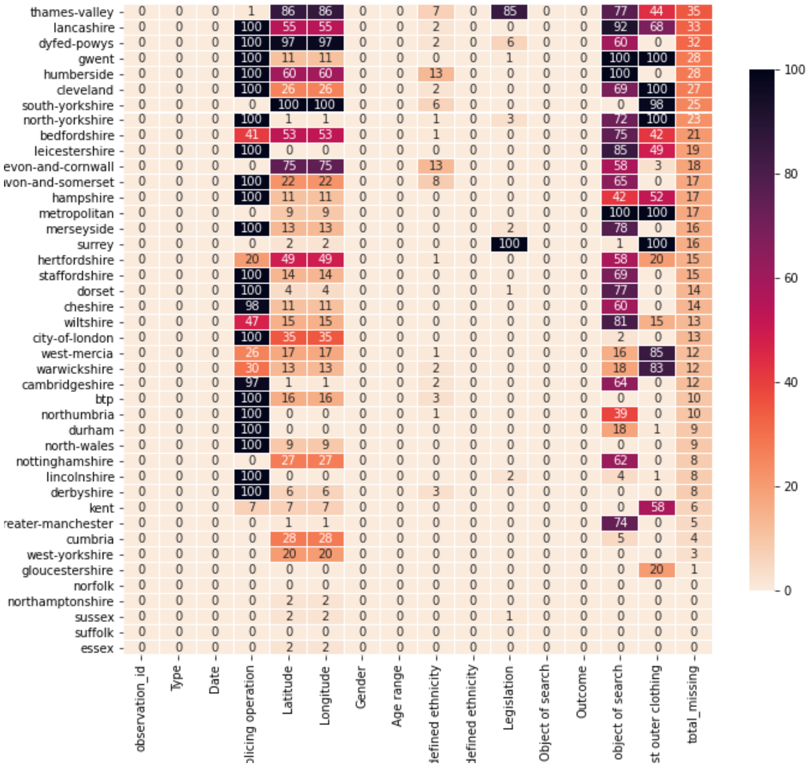
```
RangeIndex: 660611 entries, 0 to 660610
Data columns (total 16 columns):
 #   Column                                     Non-Null Count   Dtype  
---  ------                                     --------------   -----  
 0   observation_id                             660611 non-null  object 
 1   Type                                       660611 non-null  object 
 2   Date                                       660611 non-null  object 
 3   Part of a policing operation               587047 non-null  object 
 4   Latitude                                   548295 non-null  float64
 5   Longitude                                  548295 non-null  float64
 6   Gender                                     660611 non-null  object 
 7   Age range                                  660611 non-null  object 
 8   Self-defined ethnicity                     655037 non-null  object 
 9   Officer-defined ethnicity                  660611 non-null  object 
 10  Legislation                                632671 non-null  object 
 11  Object of search                           660611 non-null  object 
 12  Outcome                                    660611 non-null  object 
 13  Outcome linked to object of search         187511 non-null  object 
 14  Removal of more than just outer clothing   234062 non-null  object 
 15  station                                    660611 non-null  object 
dtypes: float64(2), object(14)
memory usage: 80.6+ MB
```

Annex 1. Original dataset



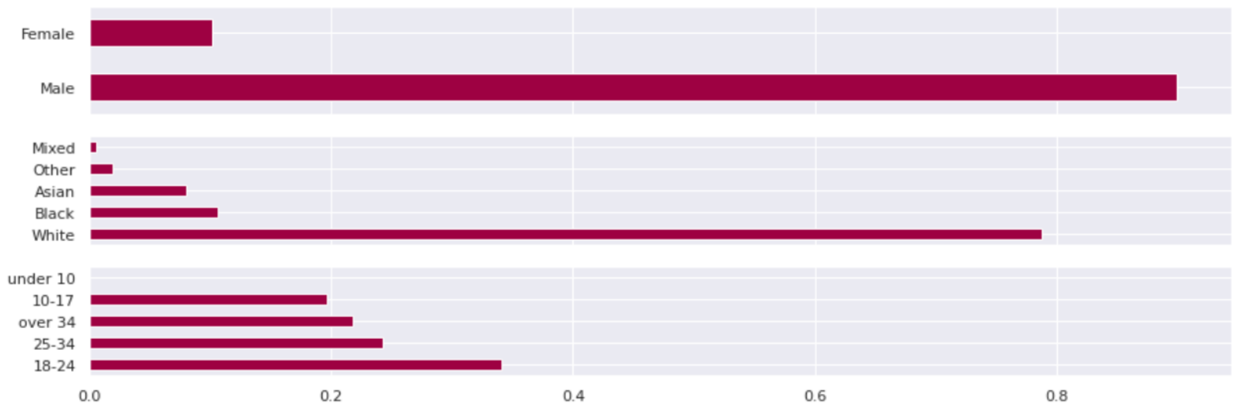
Annex. 1 Original dataset

Annex. 2 All missing data - yellow stripes

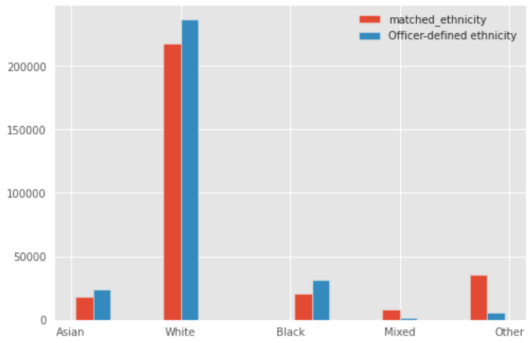


	observation_id	Type	Date	olicing operation	Latitude	Longitude	Gender	Age range	defined ethnicity	defined ethnicity	Legislation	Object of search	Outcome	object of search	st outer clothing	total_missing
thames-valley	0	0	0	1	86	86	0	0	7	0	85	0	0	77	44	35
lancashire	0	0	0	100	55	55	0	0	2	0	0	0	0	92	68	33
dyfed-powys	0	0	0	100	97	97	0	0	2	0	6	0	0	60	0	32
gwent	0	0	0	100	11	11	0	0	0	0	1	0	0	100	100	28
humberside	0	0	0	100	60	60	0	0	13	0	0	0	0	100	0	28
cleveland	0	0	0	100	26	26	0	0	2	0	0	0	0	69	100	27
south-yorkshire	0	0	0	0	100	100	0	0	6	0	0	0	0	0	98	25
north-yorkshire	0	0	0	100	1	1	0	0	1	0	3	0	0	72	100	23
bedfordshire	0	0	0	41	53	53	0	0	1	0	0	0	0	75	42	21
leicestershire	0	0	0	100	0	0	0	0	0	0	0	0	0	85	49	19
evon-and-cornwall	0	0	0	0	75	75	0	0	13	0	0	0	0	58	3	18
ıvon-and-somerset	0	0	0	100	22	22	0	0	8	0	0	0	0	65	0	17
hampshire	0	0	0	100	11	11	0	0	0	0	0	0	0	42	52	17
metropolitan	0	0	0	0	9	9	0	0	0	0	0	0	0	100	100	17
merseyside	0	0	0	100	13	13	0	0	0	0	2	0	0	78	0	16
surrey	0	0	0	0	2	2	0	0	0	0	100	0	0	1	100	16
hertfordshire	0	0	0	20	49	49	0	0	1	0	0	0	0	58	20	15
staffordshire	0	0	0	100	14	14	0	0	0	0	0	0	0	69	0	15
dorset	0	0	0	100	4	4	0	0	0	0	1	0	0	77	0	14
cheshire	0	0	0	98	11	11	0	0	0	0	0	0	0	60	0	14
wiltshire	0	0	0	47	15	15	0	0	0	0	0	0	0	81	15	13
city-of-london	0	0	0	100	35	35	0	0	0	0	0	0	0	2	0	13
west-mercia	0	0	0	26	17	17	0	0	1	0	0	0	0	16	85	12
warwickshire	0	0	0	30	13	13	0	0	2	0	0	0	0	18	83	12
cambridgeshire	0	0	0	97	1	1	0	0	2	0	0	0	0	64	0	12
btp	0	0	0	100	16	16	0	0	3	0	0	0	0	0	0	10
northumbria	0	0	0	100	0	0	0	0	1	0	0	0	0	39	0	10
durham	0	0	0	100	0	0	0	0	0	0	0	0	0	18	1	9
north-wales	0	0	0	100	9	9	0	0	0	0	0	0	0	0	0	9
nottinghamshire	0	0	0	0	27	27	0	0	0	0	0	0	0	62	0	8
lincolnshire	0	0	0	100	0	0	0	0	0	0	2	0	0	4	1	8
derbyshire	0	0	0	100	6	6	0	0	3	0	0	0	0	0	0	8
kent	0	0	0	7	7	7	0	0	0	0	0	0	0	0	58	6
reater-manchester	0	0	0	0	1	1	0	0	0	0	0	0	0	74	0	5
cumbria	0	0	0	0	28	28	0	0	0	0	0	0	0	5	0	4
west-yorkshire	0	0	0	0	20	20	0	0	0	0	0	0	0	0	0	3
gloucestershire	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	1
norfolk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
northamptonshire	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0
sussex	0	0	0	0	2	2	0	0	0	0	1	0	0	0	0	0
suffolk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
essex	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0

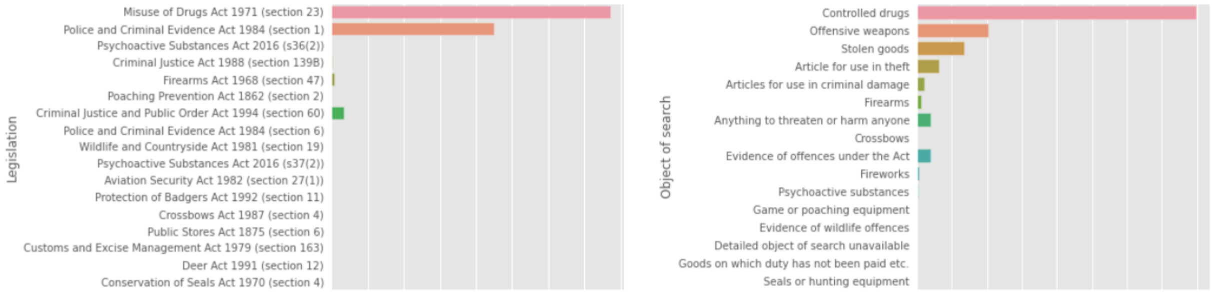
Annex. 3 % of missing data points per station and feature



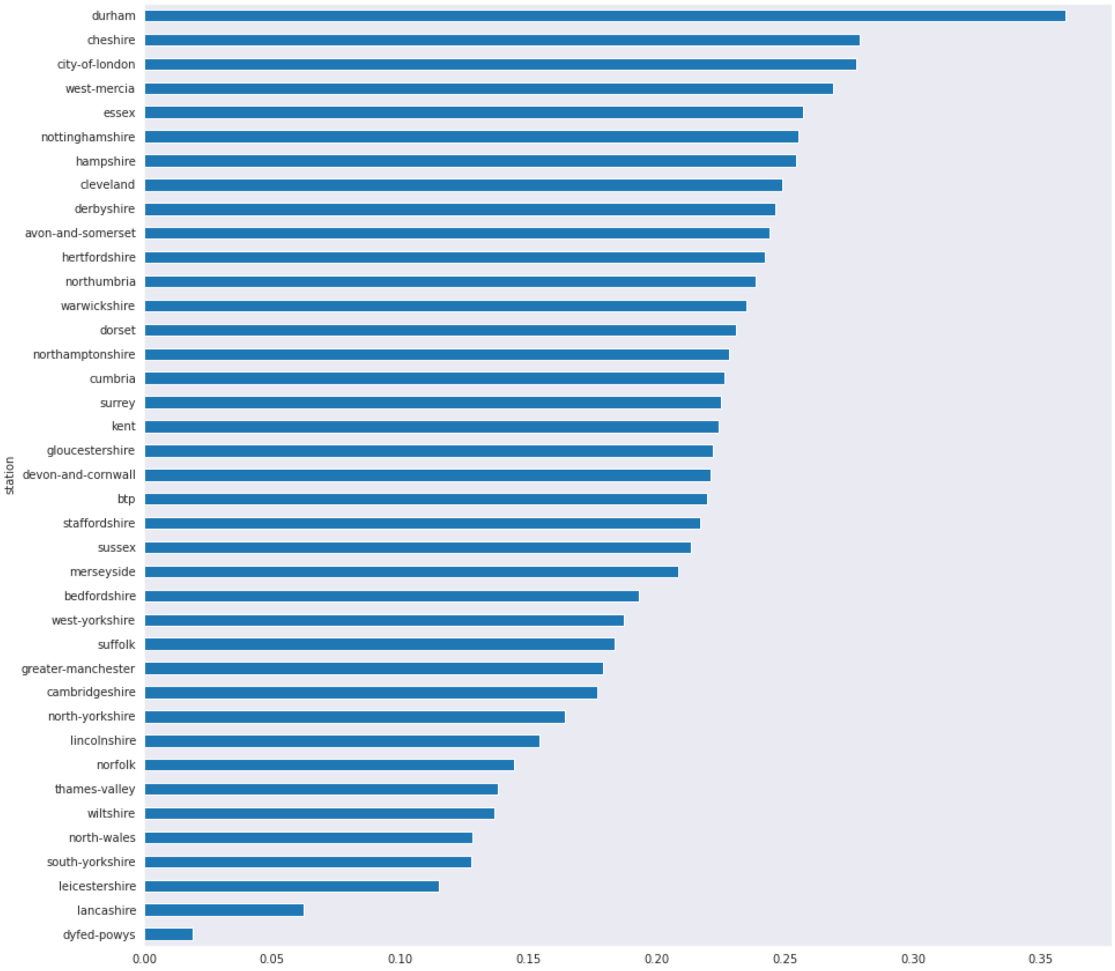
Annex 4. % of stop and search per Gender, Officer- defined ethnicity and Age range



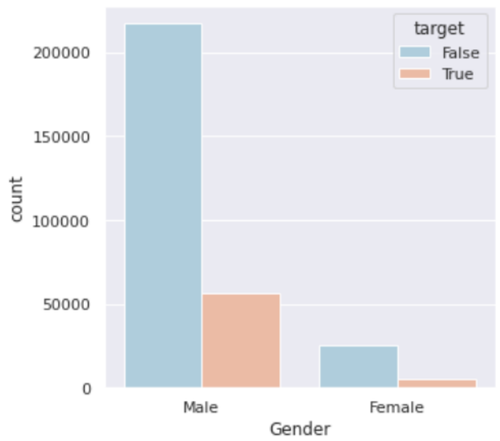
Annex 5. Match between self-defined and police-defined ethnicity



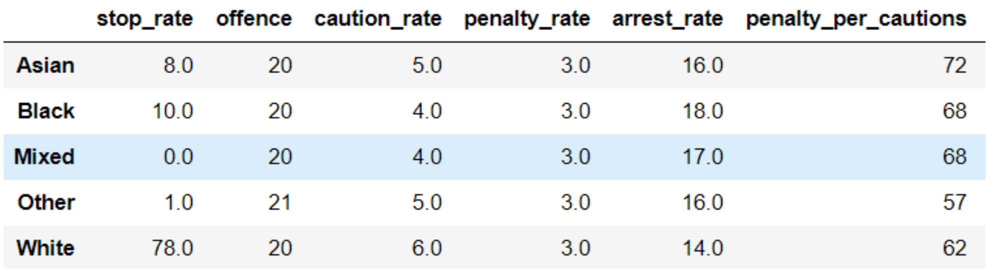
Annex 6. Values and values counts of ‘Object of search’ and ‘Legislation’



Annex 7. Distribution of discovery rate per station

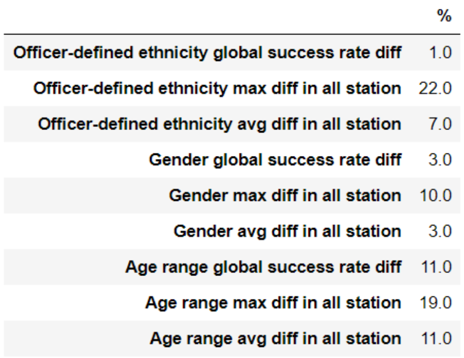


Annex 8. Global Discovery rate per gender



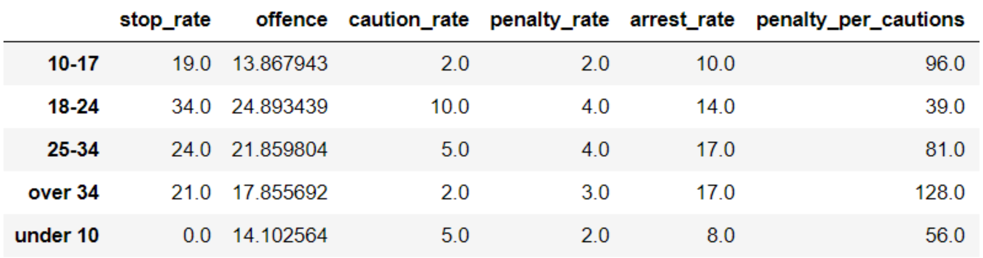
	stop_rate	offence	caution_rate	penalty_rate	arrest_rate	penalty_per_cautions
Asian	8.0	20	5.0	3.0	16.0	72
Black	10.0	20	4.0	3.0	18.0	68
Mixed	0.0	20	4.0	3.0	17.0	68
Other	1.0	21	5.0	3.0	16.0	57
White	78.0	20	6.0	3.0	14.0	62

Annexe 9.   Officer-Defined ethnic groups success rate and several other offences rates.



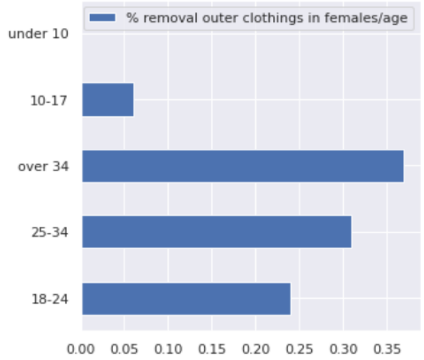
	%
Officer-defined ethnicity global success rate diff	1.0
Officer-defined ethnicity max diff in all station	22.0
Officer-defined ethnicity avg diff in all station	7.0
Gender global success rate diff	3.0
Gender max diff in all station	10.0
Gender avg diff in all station	3.0
Age range global success rate diff	11.0
Age range max diff in all station	19.0
Age range avg diff in all station	11.0

Annex 10. Nationwide maximum and average differences per sub groups. These are some model and business performance compilation of metrics that we should monitor to assess the model.



	stop_rate	offence	caution_rate	penalty_rate	arrest_rate	penalty_per_cautions
10-17	19.0	13.867943	2.0	2.0	10.0	96.0
18-24	34.0	24.893439	10.0	4.0	14.0	39.0
25-34	24.0	21.859804	5.0	4.0	17.0	81.0
over 34	21.0	17.855692	2.0	3.0	17.0	128.0
under 10	0.0	14.102564	5.0	2.0	8.0	56.0

Annex 11. Age range groups success rate and several other offences rate.



Annex 12. % removal of outer clothing asked to women per age range

Model technical analysis

The model was built as a binary classification system that outputs true or false values for stop and search. When training or fitting the model with new data the positive observations are the ones where ‘Outcome linked to object of search’ is True and where ‘Outcome’ is one of the following:

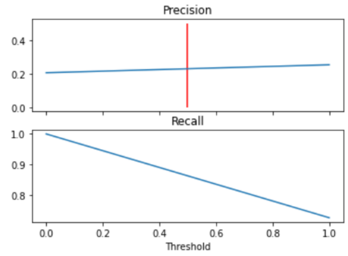
- Local resolution
- Community resolution
- Offender given drugs possession warning
- Khat or Cannabis warning
- Caution (simple or conditional)
- Offender given penalty notice
- Arrest
- Penalty Notice for Disorder
- Suspected psychoactive substances seized - No further action
- Summons / charged by post

- Article found - Detailed outcome unavailable
- Offender cautioned
- Suspect arrested
- Suspect summoned to court

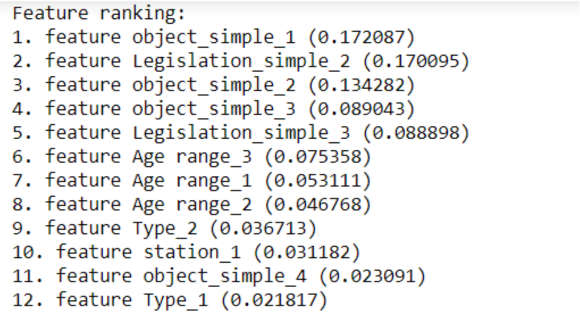
The technical specifications for the ensemble random forest model are as follows:

Impurity measure = ‘gini’;
maximum depth = 15,
minimum number of samples required to split an internal node =1,
number of trees= 100;
maximum leaf nodes=10,
maximum number of features when looking for best split = square root of number of features,
bootstrap samples = True,
out-of-bag samples = False,
class_weight="balanced" (adjust weights inversely proportional to class frequencies in the input data)

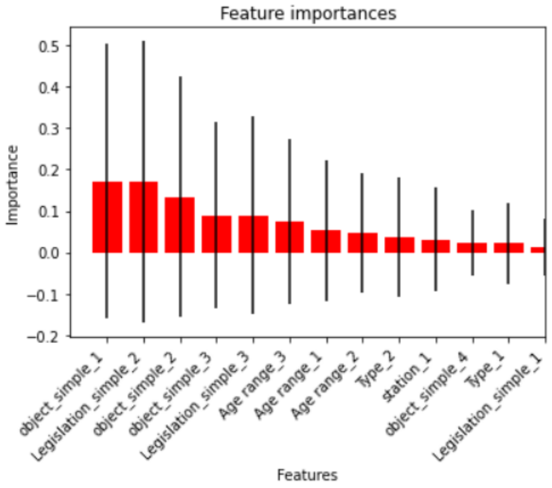
The balance between global precision and recall on the out-of-sample testing data where the decision threshold was decided on the 0.5 probability can be visualized:



Model features ranking and importance:



```
Feature ranking:
1. feature object_simple_1 (0.172087)
2. feature Legislation_simple_2 (0.170095)
3. feature object_simple_2 (0.134282)
4. feature object_simple_3 (0.089043)
5. feature Legislation_simple_3 (0.088898)
6. feature Age range_3 (0.075358)
7. feature Age range_1 (0.053111)
8. feature Age range_2 (0.046768)
9. feature Type_2 (0.036713)
10. feature station_1 (0.031182)
11. feature object_simple_4 (0.023091)
12. feature Type_1 (0.021817)
```


Feature importances
Importance
object_simple_1
Legislation_simple_2
object_simple_2
object_simple_3
Legislation_simple_3
Age range_3
Age range_1
Age range_2
Type_2
station_1
object_simple_4
Type_1
Legislation_simple_1
Features