Stop and Search Policy - The Data Science Approach

Client Requirements

Summary

This report aims to help the United Kingdom Department of Police with their stop and search policy for every police station through data science. Our analysis will help to identify and quantify some of the accusations the Department has been subject to, namely issues of discrimination against minority groups in performing searches without probable cause and asking suspects to remove articles of clothing for search.

After analysing the collected data, in addition to evaluating the credibility of current claims and helping to defend the Department's public image, we will also provide some recommendations for the Department to improve their performance on some of the 42 stations available. These studies will allow the development of a platform that can be used by each Officer to validate a stop and search request through an API. This approach will allow the Department to monitor a greater number of occurrences in real time and quickly be able to react to new situations that may occur.

Requirements clarifications

Our API will be running for a year and will only validate a request for search if the predicted probability (of successful search) is higher than 10%. In practice, our goal will be to keep model's Recall close to 100% (high discovery rate) while maximizing Precision (overall ability to detect offences), ensuring uniformity of decisions across stations, search objectives and ethnicities.

Regarding discrimination claims, we will consider a minority group to be any combinations of station/ethncity/gender with a minimum of 30 occurrence in the initial dataset provided.

We will consider 6 KPIs to evaluate discrimination: KPIs to measure each station's performance and KPIs to evaluate the Department global behaviour. For sub-groups within the same station, we will assume a 5% threshold difference to station's average in search success rate to assess dicrimination. Between stations, a 10% threshold was considered when comparing to overall Department average and stations with null success rate were not used.

A search is considered successful if the outcome is positive and related to the object of search.
Positive outcomes are all those who confine any transgression to the law, in our case we have:

1. Arrest
2. Community resolution
3. Summons / charged by post
4. Khat or Cannabis warning
5. Caution (simple or conditional)
6. Penalty Notice for Disorder
7. Offender given drugs possession warning
8. Local resolution
9. Suspect arrested
10. Article found - Detailed outcome unavailable
11. Offender cautioned
12. Suspect summoned to court
13. Offender given penalty notice
14. Suspected psychoactive substances seized - No further action

On the other hand, negative outcomes are those who do not violate the law:

1. A no further action disposal
2. Nothing found - no further action

We detected some suspicious entries in the dataset, there are cases where the suspect did not violate the law but the 'Outcome linked to object of search' is True. Please consult annexes for more details.

Dataset Analysis

General analysis

We will start by analysing the initial dataset variables (columns, features) and all values/categories they can have. Missing data and how to overcome it in some situations will also be addressed. In another chapter we will analyse which stations have more missing data.

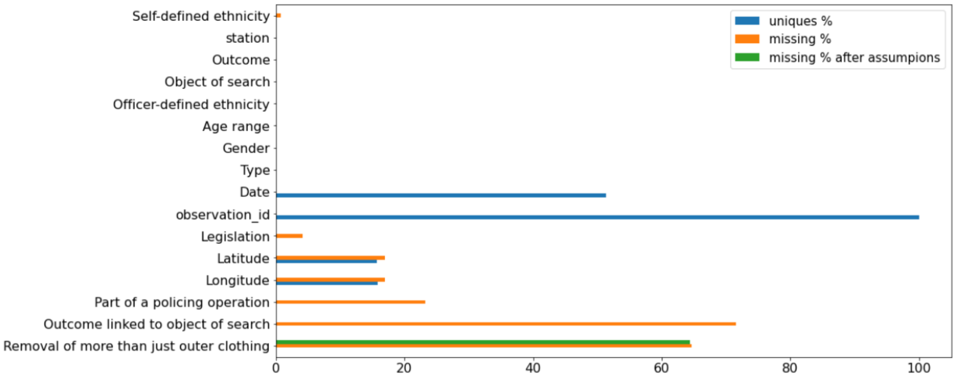
Current dataset has 16 features of which 3 will not be available for the API (Self-defined ethnicity and both Outcome related features). To overcome missing data, we assumed the following:

- Removal of more than just outer clothing
  - In a 'Vehicle search' occurrence it is very likely that the suspect will not be searched, so we replace it with False
- Outcome linked to object of search
  - We replace all missing values with False because Officers hardly miss an opportunity to fill a successful search
- Part of a policing operation
  - Given that there's only 3.2% of True reported cases, we will assume that these are very rare situations, therefore we replace missing values with False
- Latitude and Longitude
  - There are 17% of cases with missing locations. We used each station statistics mode value (or average of mode values) to replace missing data. We also noticed that station 'south-yorkshire' do not have any data so we will use the dataset average

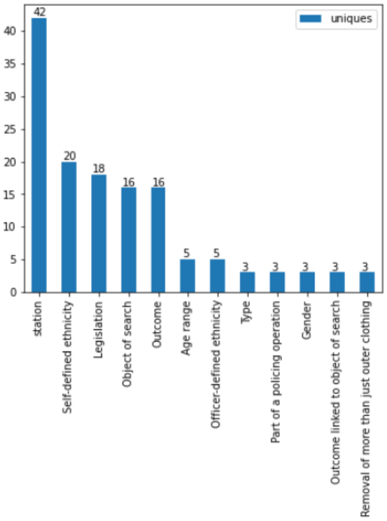
- Legislation
  - Only 4.2% of cases have missing values and we found a high correlation between 'Legislation' and 'Object of search' for all values. Therefore, we replaced any missing value in 'Legislation' with the correspondent statistics mode for equal 'Object of search' values. Another options would be to discard this feature and only use 'Object of search' instead
- Self-defined ethnicity
  - Since this feature will not be available for the API and we have other feature with similar information (Officer-defined ethnicity), we will not use this one in our analysis. We will only look for stations with a big discrepancy between these two, which may indicate some less proper behaviour

After our assumptions, we have missing values only for cloth removal.

We also analysed the amount of different values each feature can assume (unique values). There is a distinct value for observation_id in each occurrence, which is good because we don't have duplicate data. Date, Latitude and Longitude have a large percentage of unique values, this topic will be addressed later because it may have a big impact on our model performance.



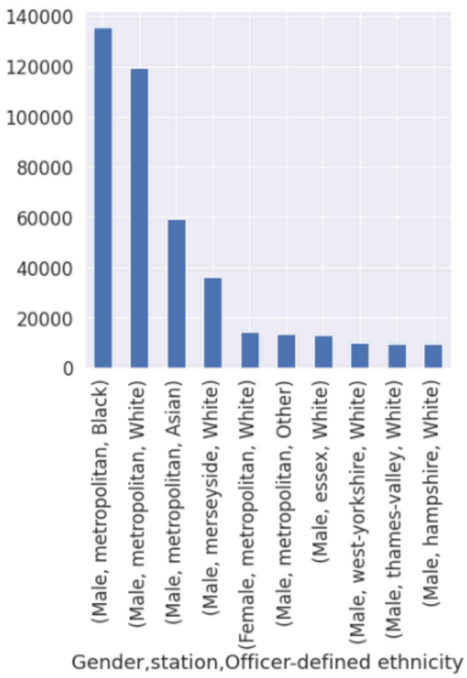
All other features have a manageable amount of unique values (including missing category).



We chose to group categories from 'Self-defined ethnicity' to match the 5 categories from 'Officer-defined ethnicity'.

Business questions analysis

In this section we will analyze the accusations that have been made to the Department regarding discrimination of certain minorities in search operations and the criteria for removing clothes.



We created groups based on common Gender + station + Officer-defined ethnicity, but from all 630 possible combinations, we only have 401 in our dataset. We will only use groups well represented and we assumed it to be more than 30 occurrences. This approach reduced our dataset to 99.8% of initial occurrences and to 243 groups, which means an average of 6 sub-groups per station.

By analysing top 10 more representative groups, we see a dominance of 'Male', 'metropolitan' and 'White', despite the largest group having 'Black' and 20.5% of cases. We also noticed some groups and stations with 0 successful searches, thus null success rates (we recommend a recheck on this).

To assess the claims of discrimination within each station we crossed information of sub-groups and search rates (KPI-1 in Annexes).

Our comparison between stations performance (KPI-2 in Annexes) shows 15 stations below global mean rate margin (not considering all 3 stations with 0 rate). Department mean rate is 20.4%.
Similarly, we analysed the claim regarding cloth removal using the same method (KPI-3 in Annexes). This time, we have sub-groups of station + Gender + Officer-defined ethnicity + Age range. Only 35 out of 42 stations presented reliable data (no missing data in all sub-groups). Results show 7 stations where at least 50% of sub-groups are being discriminated to remove clothes.

Ranking (KPI-4 in Annexes) show that sub-group 'Male'+'Asian'+'over 34' is considered as a discriminated sub-group in 54.3% of 35 stations. Overall, ´White' dominates the discrimination across all stations.

All these 4 KPIs show which stations are most likely to be contributing to current claims. However, current data does not indicate a clear discriminatory behaviour in the Department.

We also analysed the Officer performance when categorizing the suspect. In a way, we quantify ethnic discrimination by the Department's ability to perceive the suspect's ethnicity, comparing 'Self-defined ethnicity' and 'Officer-defined ethnicity' (see KPI-5 and KPI-6 in Annexes).

Conclusions and recommendations

Our first recommendation is a recheck on data from stations:

- Gwent, Humberside and Metropolitan - no data for successful occurrences
- Cleveland, Gwent, Metropolitan, North-yorkshire and Surrey - no data when Officer asks to remove cloth
- Gwent, North-yorkshire, South-yorkshire, Surrey - no data when Officer does not ask to remove cloth

These 7 stations represent 58% of occurrences so any data update may influence this report conclusions. Metropolitan station alone has 53% of total cases.

Regarding discrimination claims on stop and search policy, our analysis show some problematic stations when comparing sub-group search rates, namely Avon-and-somerset, Nottinghamshire, Leicestershire, Dyfed-powys and Wiltshire.

Current claims on discrimination when asking suspects to remove articles of clothing for search are more evident in stations Bedfordshire, Btp, Cambridgeshire, North-wales, Greater-manchester, Nottinghamshire and Suffolk. These stations reveal a discriminatory behaviour in 50% or more of their sub-groups.

Officer's accuracy on correctly defining the suspect's ethnie was evaluated at each station and globally. Results show an average successful ethny definition in 77% of cases and stations City-of-london, Bedfordshire and Btp scored below 70%. In a general perspective, 'Mixed' and 'Asian' ethnies are difficult for Officer's to recognize.

Combining all the analysis performed and KPI rankings, we can conclude in which stations we can change behaviours to improve the Department public opinion.

In future analysis we plan to analyse all these behaviors, claims and KPIs over time. This way it will be possible to see if the measures that the Department has implemented will favor public opinion or not.

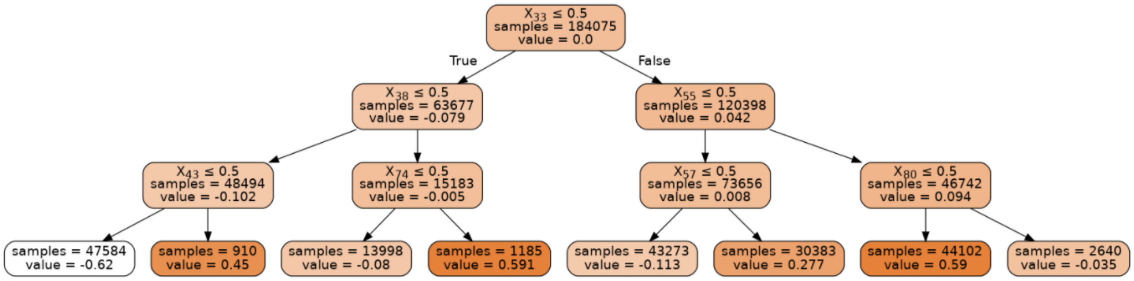
Modeling

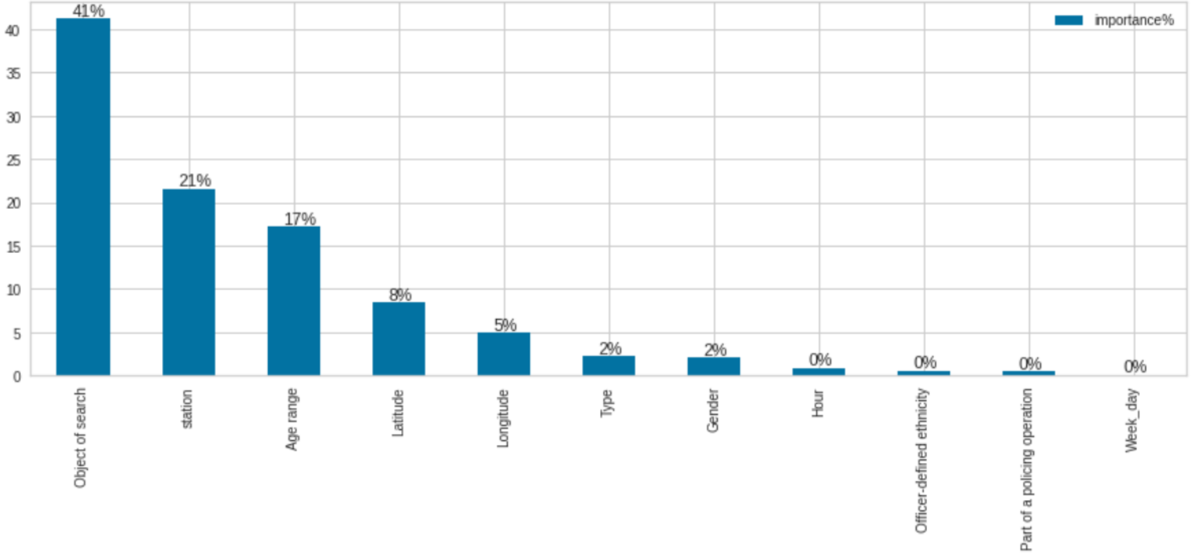
Model expected outcomes overview

Our model will predict the probability of a successful search and authorize when it's probability exceeds 10%. We found out that this 10% threshold minimizes the number of suspects we miss, in other words, we search as many wrongdoers (successful search) as possible without searching every suspect. We used the F_beta score to evaluate and choose a model, reaching a score of 86,9% with a beta equals to 5 to increase the recall's relative contribution to precision.

$$F_\beta = \frac{(1+\beta^2)\cdot(\text{precision}\cdot\text{recall})}{(\beta^2\cdot\text{precision}+\text{recall})}$$

Our model consists of an ensemble of decision trees. We can evaluate each decision tree (example below), but we extract more knowledge by looking at each feature importance.





With the current model, we can expect a great sensitivity from our API with 'object of search'. Given that data from some stations is suspect and may undergo significant changes, 'station' importance may change. Ethnicity and Gender have very little influence on our model.

Model specifications

Our first step to create a model was to clean the dataset. We dropped all features which will not be available in API requests. Next, we removed some stations from data because we suspect that there was a problem collecting data and those problems would negatively influence our model:

- Cleveland, Metropolitan and Humberside do not have successful searches
- South-yorkshire always asks suspects to remove cloth. Despite not being a valid feature for our model, we suspect this behavior may not be realistic

We also removed 'Legislation' because it provides the same information as 'Object of search'. 'Date' was used to create 2 new features, Hour and WeekDay. These new features will be used as categories in the model.

The second step to create a model was to prepare our pipeline workflow. We used 3 groups of features (see Annexes for more details):

- Categorical using One-Hot-Encoding
  - Part of a policing operation
  - Hour
  - Week_day
  - Object of search
  - Officer-defined ethnicity',
  - Age range
  - Gender
  - Type
- Categorical using Target-Encoder
  - station
- Numerical using KBins-Discretizer
  - Latitude
  - Longitude

This approach allowed us to evaluate all possible categories with a reasonable computation trade-off. Discretization in location features allows our model to look for specific locations/areas.

Our current model is using 'Gradient-Boosting-Classifier' to predict the necessary outcome (probability). Other classifiers were tested but this one stood out from the rest. Performance will be presented in the following chapters and additional parameter details in annexes.

We used 80% (train set) of the dataset to train our model and 20% (validation set) was used to evaluate/test it's performance. In order to reproduce a more realistic situation, the dataset must be ordered by 'Date' prior to splitting.

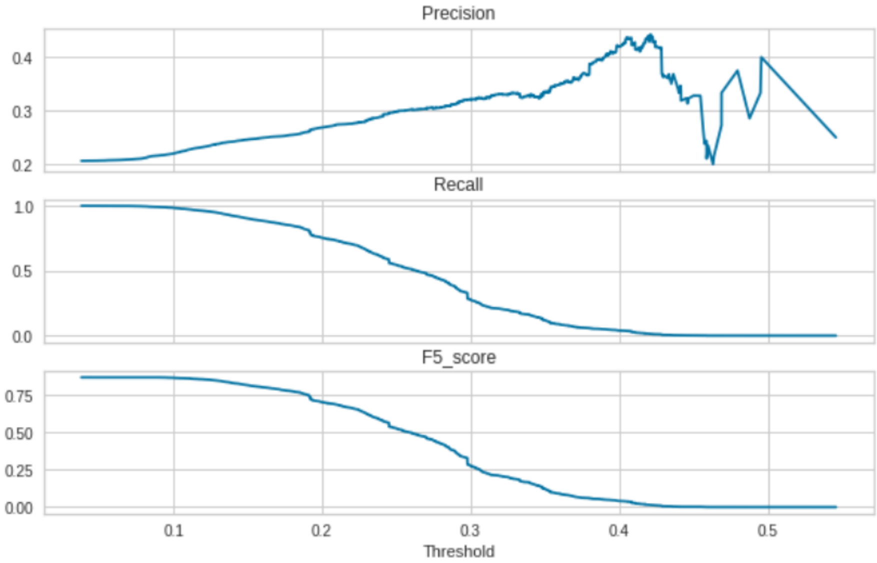
To minimize missing any wrongdoer suspect (suspect we know will be a successful search), our model will authorize any request for search (API returns true) with a probability above 10%.

Analysis of expected outcomes based on training set

As requested, we prepared our model to increase the discovery rate while keeping a high overall ability to detect offenses, leveling all ethnicities. Discovery rate will be measured as the percentage of wrongdoer suspects we are able to correctly identify (recall), whereas the overall ability to detect offenses will be measured by the percentage of occurrences correctly predicted by our model (precision). Presented results were obtained using the validation set and our model pipeline.

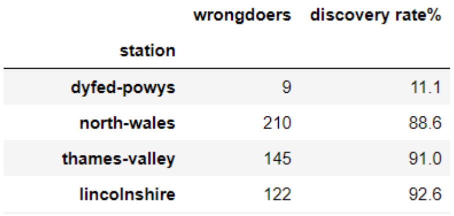
Current model has a recall of 98.5% and 22% precision, which means it's only correctly predicting 22% of occurrences but it hardly misses any opportunity to search a wrongdoer suspect (only 1.5%). Therefore, we have a F5 score of 86.9%.

Raising the threshold to more than 10% would result in a poorer model performance, as shown by these graphs.



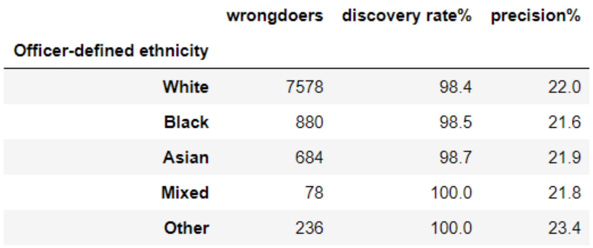
When looking at the performance of our model across all stations, we obtain a successful search rate of 22% (successful searches from cases predicted as true) which is a minor improvement from previous 20.4% (from KPI-2 analysis).

We identified which stations from our validation set have a lower discovery rate (positive prediction of wrongdoers). Dyfed-powys stood out with a poor performance probably due to lack of data.



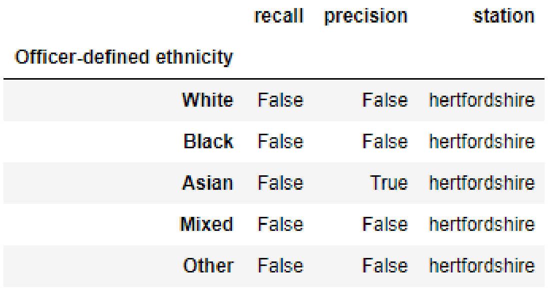
	wrongdoers	discovery rate%
station		
dyfed-powys	9	11.1
north-wales	210	88.6
thames-valley	145	91.0
lincolnshire	122	92.6

When looking at the average discovery rate and average precision between ethnicities we found a balanced distribution, which indicates that generally our model does not discriminate any ethnicity.

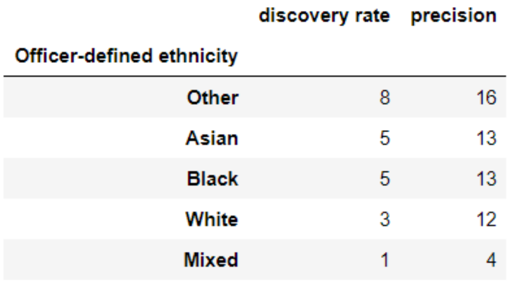


	wrongdoers	discovery rate%	precision%
Officer-defined ethnicity			
White	7578	98.4	22.0
Black	880	98.5	21.6
Asian	684	98.7	21.9
Mixed	78	100.0	21.8
Other	236	100.0	23.4

Discovery rate and precision were evaluated for each ethnic group in each station (33 in the validation set), and compared to the global average within a 10% margin.
As an example, in 'hertfordshire' station only precision for 'Asians' perform below the Department's average value. Results show that 'Other' group have a discovery rate below average margin in 8 stations and precision in 16 stations is also below average margin.

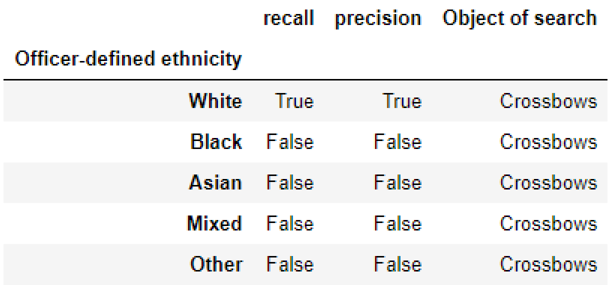


	recall	precision	station
Officer-defined ethnicity			
White	False	False	hertfordshire
Black	False	False	hertfordshire
Asian	False	True	hertfordshire
Mixed	False	False	hertfordshire
Other	False	False	hertfordshire

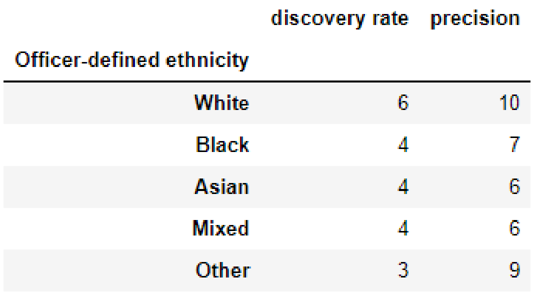


	discovery rate	precision
Officer-defined ethnicity		
Other	8	16
Asian	5	13
Black	5	13
White	3	12
Mixed	1	4

Similarly, 'Object of search' was evaluated for each ethnic group. As an example, searching for 'Crossbows' shows recall and precision below the Department's average margin in a 'White' suspect. Generally from 14 'Object of search' categories, 6 categories under-perform in discovery rate (recall) and 10 categories precision are below mean margin.



	recall	precision	Object of search
Officer-defined ethnicity			
White	True	True	Crossbows
Black	False	False	Crossbows
Asian	False	False	Crossbows
Mixed	False	False	Crossbows
Other	False	False	Crossbows



	discovery rate	precision
Officer-defined ethnicity		
White	6	10
Black	4	7
Asian	4	6
Mixed	4	6
Other	3	9

Alternatives considered

In the development phase, other classifiers were evaluated (see Annexes for details). A similar performance is observed for ‘GradientBoosting’, ‘DecisionTree’, ‘RandomForest’, ‘MLP’ and ‘AdaBoost’.

We will give preference to 'GradientBoosting' (explanation) because:

- Performs better than ‘DecisionTree’ and ‘RandomForest’ for unbalanced data (our current scenario). This behaviour is a consequence of weighing each decision tree differently.
- Has better explainability than ‘MLP’. ‘MLP’ is a neural network classifier which will create and use additional variables/features instead of our own.
- It’s more flexible than ‘AdaBoost’. ‘GradientBoosting’ it’s a generalization of ‘AdaBoost’ and more suitable for noisy data.

There are several parameters to tune in ‘GradientBoosting’ but we focused only on (27 combinations):

- max_depth: [3, 4, 5]
- learning_rate: [0.01, 0.1, 1]
- n_estimators: [100, 150, 200]

Results (see Annexes) show very little improvements so we decided to keep the default parameters from scikit learn (max_depth=3, learning_rate=0.1 and n_estimators=100).

Known issues and risks

The main risk when using this model daily throughout the Department will be when predicting a request for new and/or insufficient data (i.e. new feature, new category or a station that did not provide sufficient valid data in the development of this model). The change in the conjuncture (external factors) compared to the period in which the training data was collected, may make this model obsolete (i.e. extreme climatic, political or social conditions; one-off events with large groups of people, such as riots, sporting events, etc.).

Current model requires careful tuning of hyperparameters to avoid overfitting with too many trees (n_estimators). One way to address this issue is to update the model regularly.

Features importances (shown above) generally describe the decisions our model makes with available features. However, it’s possible that any tree decision might be introducing bias and/or discriminating a minor portion of our dataset.

‘GradientBoosting’ is sensitive to outliers because each tree is built on previous trees' residuals/errors. Outliers and errors in observed data labels may have a strong influence in the current model (reference).

Model Deployment

Deployment specifications

Our model was deployed in Heroku to provide an API and store requests in a Postgres database.

The API have 2 endpoints (capabilities):

- should_search/
    - This endpoint handles information about the stop and search request
    - When submitting a request, information must be formatted according to specification, example:

      ```
      {
      "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”
      }
      ```
    - “Observation_id” and predicted outcome are stored in Database
    - With the exception of “Observation_id”, we consider that all information provided is confidential data and will therefore not be stored in the database
- search_result/
    - This endpoint handles information about the actual outcome of a stop and search request
    - When submitting a request, information must be formatted according to specification, example:

      ```
      {
      "observation_id": “string”,
      "outcome": boolean
      }
      ```
    - “outcome” is stored in the Database, updating the “observation_id” entry

Known issues and risks

The main issue while using this API is to submit a request with an unknown value in any feature, for example:

- a request from a station that was not used in the model's training
- a request without "Age range"
- a request with an extra feature such as "weather": "sun"

To overcome these possible issues, current API will only use features and categories that were considered in initial data. When necessary, default values will be used and extra features will not be considered.

Current default values are:

- Type: most frequent value in training
  - Person search
- Part of a policing operation: most frequent value in training
  - False
- Gender: most frequent value in training
  - Male
- Age range: most frequent value in training
  - 18-24
- Officer-defined ethnicity: most frequent value in training
  - White
- Object of search: most frequent value in training
  - Controlled drugs
- station: most frequent value in training
  - merseyside
- Latitude: mean value in training
  - 52.549171
- Longitude: mean value in training
  - -1.514060
- observation_id:
  - 00000000-0000-0000-0000-000000000000
- Date: converted to Hour and Week_day most frequent values in training
  - Hour = 14
  - Week_day = 3

Another source of risks and issues are the Heroku platform and PostgreSQL DB we will be working with. Heroku with a PostgreSQL DB has its own limits of which we can highlight:

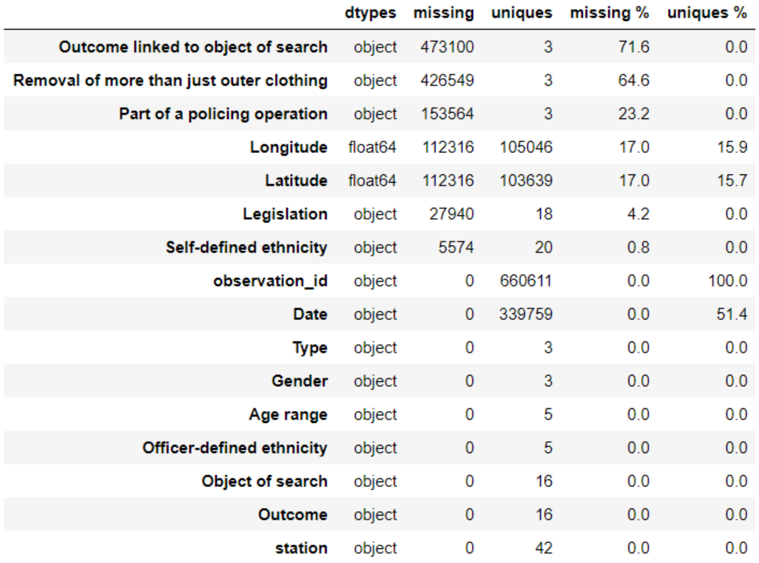
- Stores only most recent 1500 lines of log history
- Limit of 4500 API requests per hour
- Dataclips limit to 100,000 rows

We also emphasize that the availability of the API with all information control measures and data confidentiality will be dependent on the good faith and correct functioning of the platform.

Annexes

Dataset technical analysis

Initial dataset have 16 features (columns) and 660611 cases (rows) with the following types, missing rows and unique values:

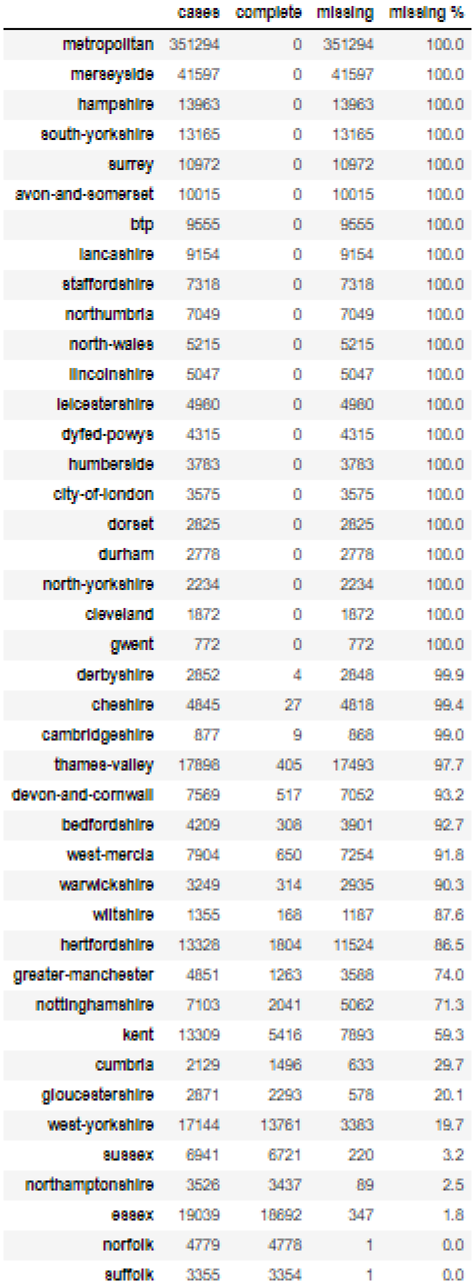


	dtypes	missing	uniques	missing %	uniques %
Outcome linked to object of search	object	473100	3	71.6	0.0
Removal of more than just outer clothing	object	426549	3	64.6	0.0
Part of a policing operation	object	153564	3	23.2	0.0
Longitude	float64	112316	105046	17.0	15.9
Latitude	float64	112316	103639	17.0	15.7
Legislation	object	27940	18	4.2	0.0
Self-defined ethnicity	object	5574	20	0.8	0.0
observation_id	object	0	660611	0.0	100.0
Date	object	0	339759	0.0	51.4
Type	object	0	3	0.0	0.0
Gender	object	0	3	0.0	0.0
Age range	object	0	5	0.0	0.0
Officer-defined ethnicity	object	0	5	0.0	0.0
Object of search	object	0	16	0.0	0.0
Outcome	object	0	16	0.0	0.0
station	object	0	42	0.0	0.0

Missing values will lead to a poor performance in our model. Analysing the missing data at each station will allow us to know where to focus efforts to improve our model.

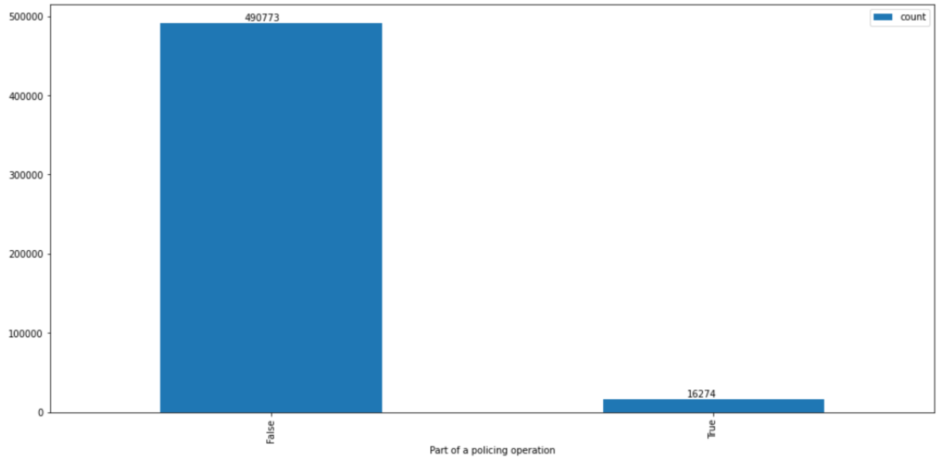
There are 42 stations and we can see that 21 stations (50%) have missing values in all cases (complete=0).

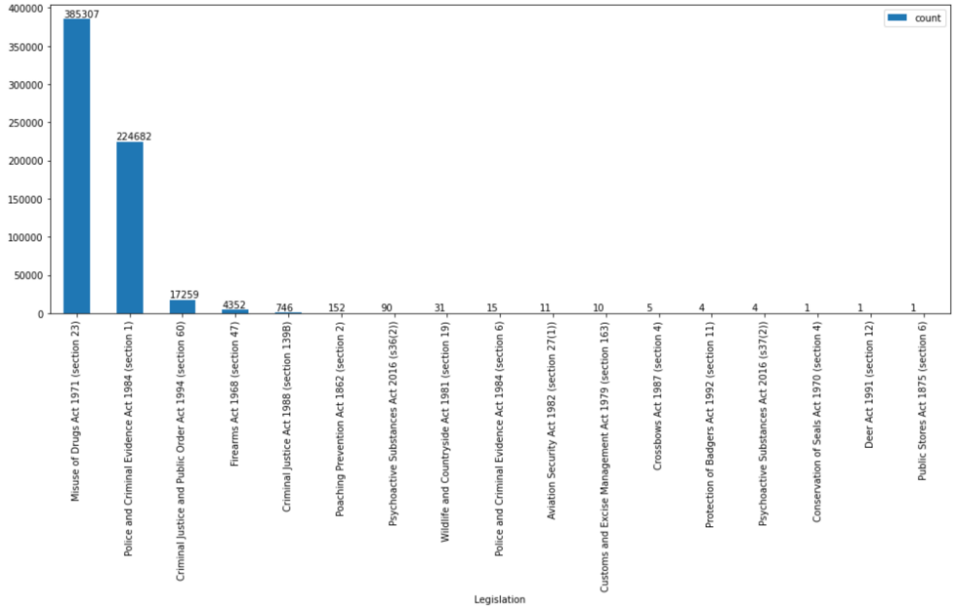
There are only 10.2% cases without any missing value. If we contact metropolitan station and receive all missing data, we would improve to 63.4% cases.

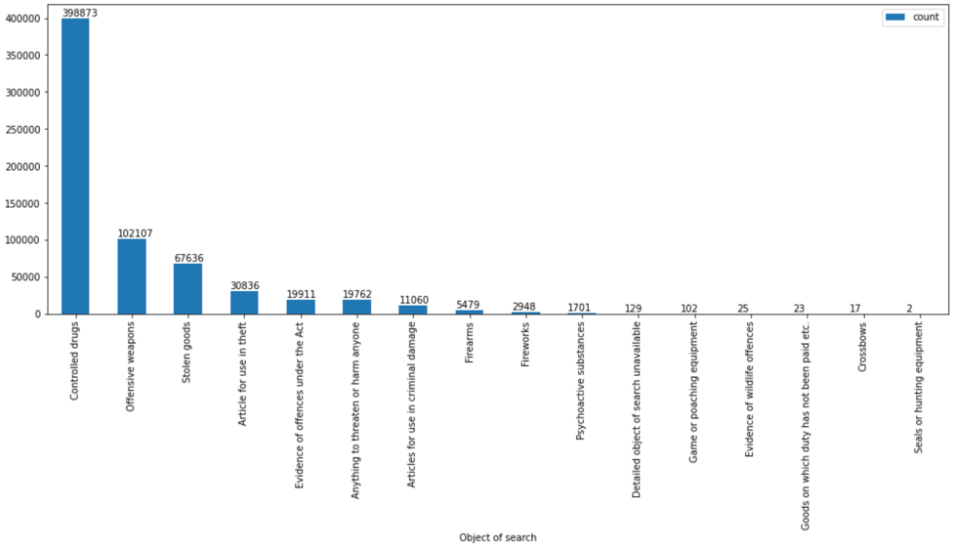


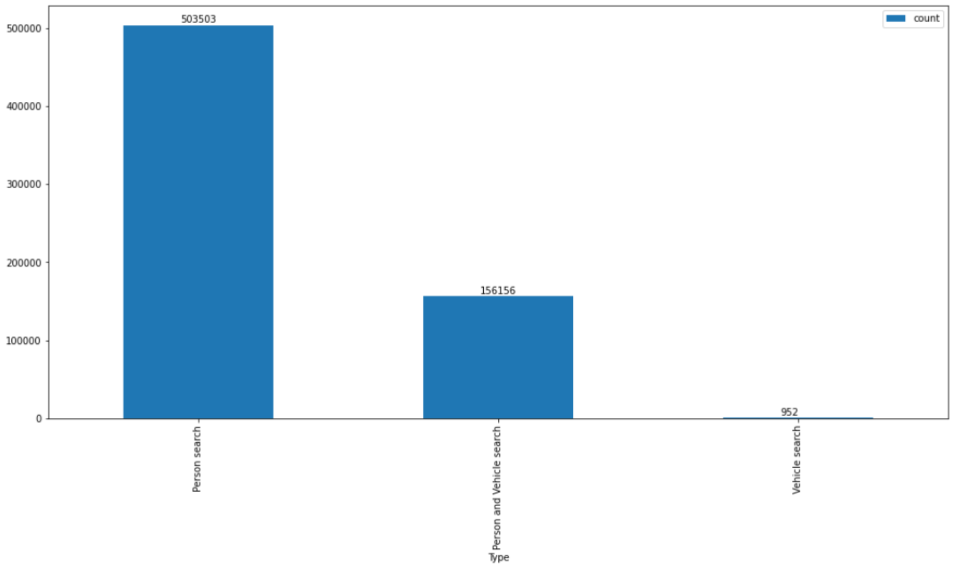
	cases	complete	missing	missing %
metropolitan	351294	0	351294	100.0
merseyside	41597	0	41597	100.0
hampshire	13963	0	13963	100.0
south-yorkshire	13165	0	13165	100.0
surrey	10972	0	10972	100.0
avon-and-somerset	10015	0	10015	100.0
btp	9555	0	9555	100.0
lancashire	9154	0	9154	100.0
staffordshire	7318	0	7318	100.0
northumbria	7049	0	7049	100.0
north-wales	5215	0	5215	100.0
lincolnshire	5047	0	5047	100.0
leicestershire	4980	0	4980	100.0
dyfed-powys	4315	0	4315	100.0
humberside	3783	0	3783	100.0
city-of-london	3575	0	3575	100.0
dorset	2825	0	2825	100.0
durham	2778	0	2778	100.0
north-yorkshire	2234	0	2234	100.0
cleveland	1872	0	1872	100.0
gwent	772	0	772	100.0
derbyshire	2852	4	2848	99.9
cheshire	4845	27	4818	99.4
cambridgeshire	877	9	868	99.0
thames-valley	17898	405	17493	97.7
devon-and-cornwall	7569	517	7052	93.2
bedfordshire	4209	308	3901	92.7
west-mercia	7904	650	7254	91.8
warwickshire	3249	314	2935	90.3
wiltshire	1355	168	1187	87.6
hertfordshire	13328	1804	11524	86.5
greater-manchester	4851	1263	3588	74.0
nottinghamshire	7103	2041	5062	71.3
kent	13309	5416	7893	59.3
cumbria	2129	1496	633	29.7
gloucestershire	2871	2293	578	20.1
west-yorkshire	17144	13761	3383	19.7
sussex	6941	6721	220	3.2
northamptonshire	3526	3437	89	2.5
essex	19039	18692	347	1.8
norfolk	4779	4778	1	0.0
suffolk	3355	3354	1	0.0

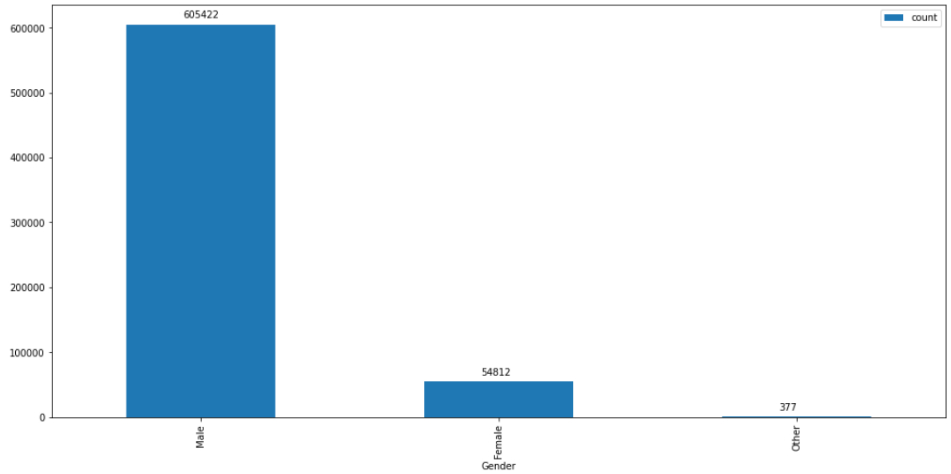
Next pages show the distribution of all other categorical features.

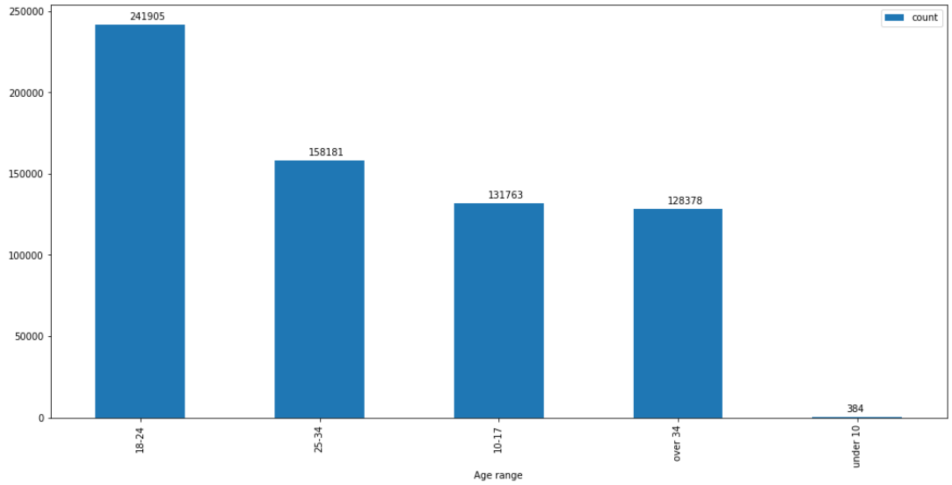


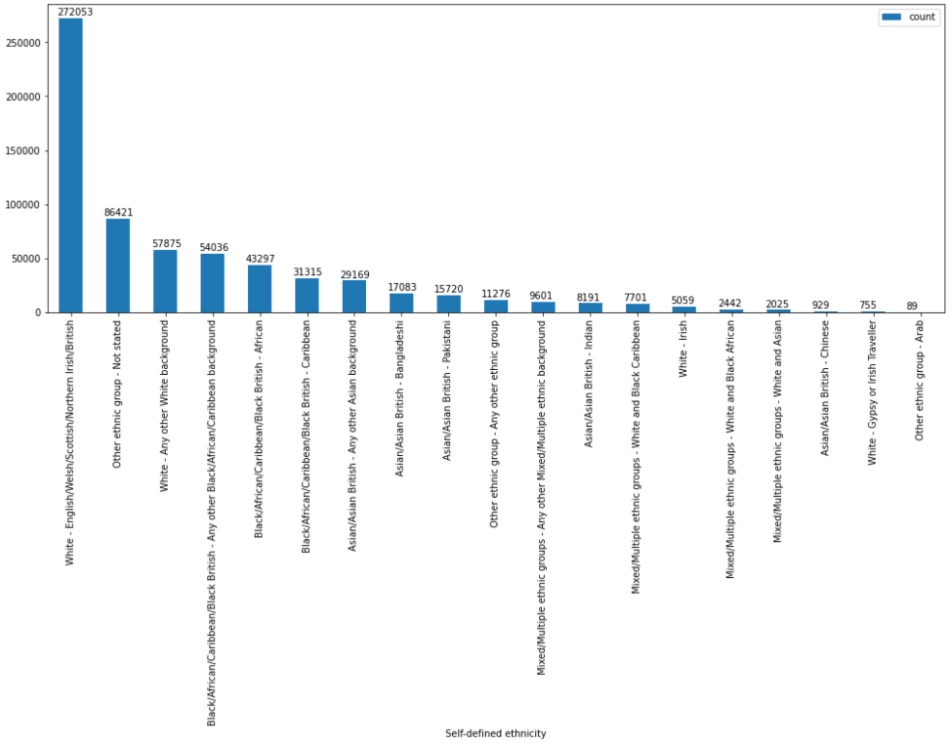


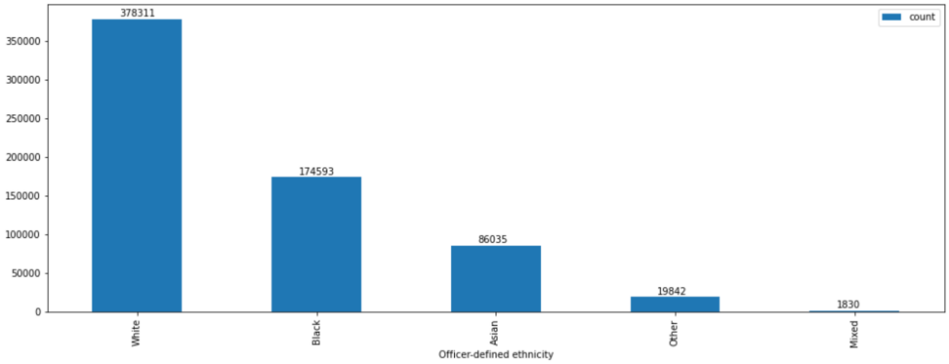


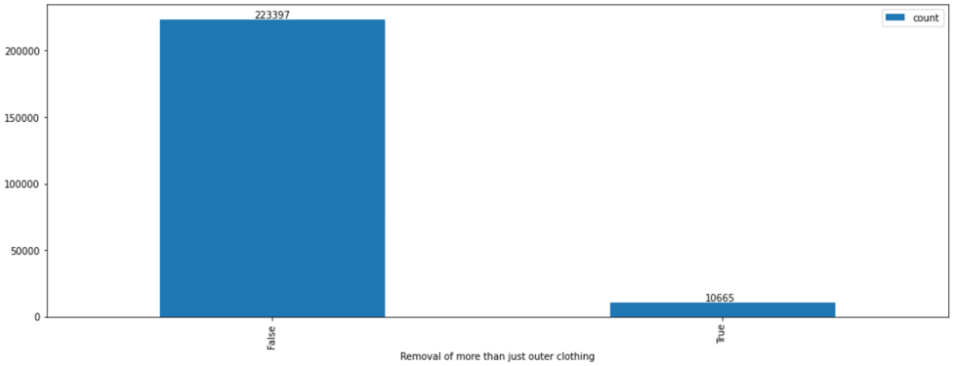

count
503503
156156
952
500000
400000
300000
200000
100000
0
Person search
Person and Vehicle search
Vehicle search
Type

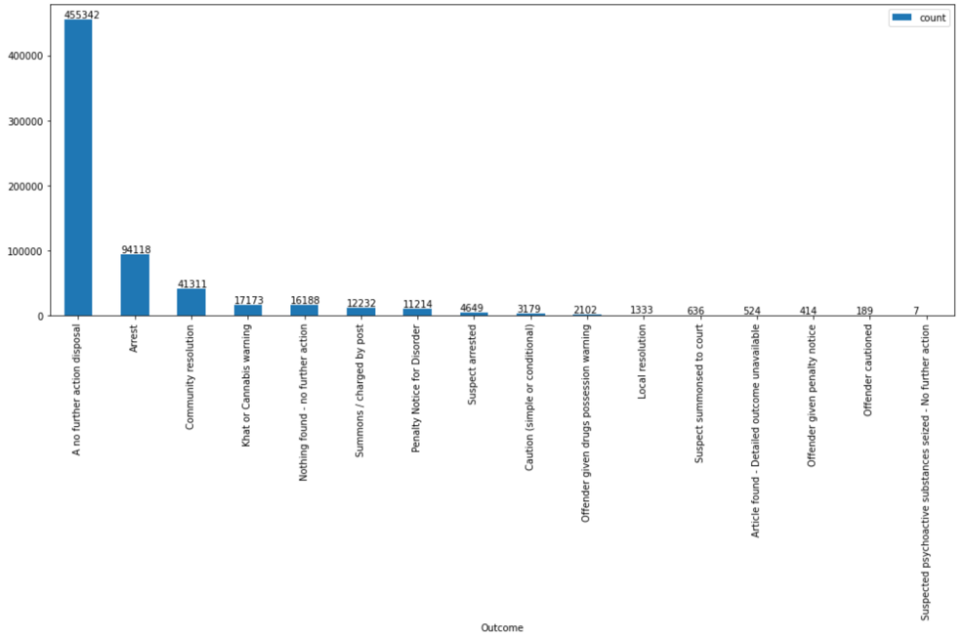

count
605422
54812
377
600000
500000
400000
300000
200000
100000
0
Male
Female
Other
Gender

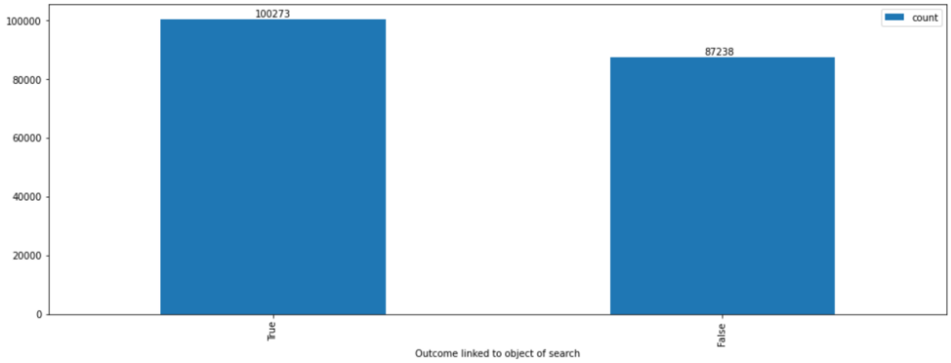

count
241905
158181
131763
128378
384
250000
200000
150000
100000
50000
0
18-24
25-34
10-17
over 34
under 10
Age range


count
0
50000
100000
150000
200000
250000
272053
86421
57875
54036
43297
31315
29169
17083
15720
11276
9601
8191
7701
5059
2442
2025
929
755
89
White - English/Welsh/Scottish/Northern Irish/British
Other ethnic group - Not stated
White - Any other White background
Black/African/Caribbean/Black British - Any other Black/African/Caribbean background
Black/African/Caribbean/Black British - African
Black/African/Caribbean/Black British - Caribbean
Asian/Asian British - Any other Asian background
Asian/Asian British - Bangladeshi
Asian/Asian British - Pakistani
Other ethnic group - Any other ethnic group
Mixed/Multiple ethnic groups - Any other Mixed/Multiple ethnic background
Asian/Asian British - Indian
Mixed/Multiple ethnic groups - White and Black Caribbean
White - Irish
Mixed/Multiple ethnic groups - White and Black African
Mixed/Multiple ethnic groups - White and Asian
Asian/Asian British - Chinese
White - Gypsy or Irish Traveller
Other ethnic group - Arab
Self-defined ethnicity


count
0
50000
100000
150000
200000
250000
300000
350000
378311
174593
86035
19842
1830
White
Black
Asian
Other
Mixed
Officer-defined ethnicity

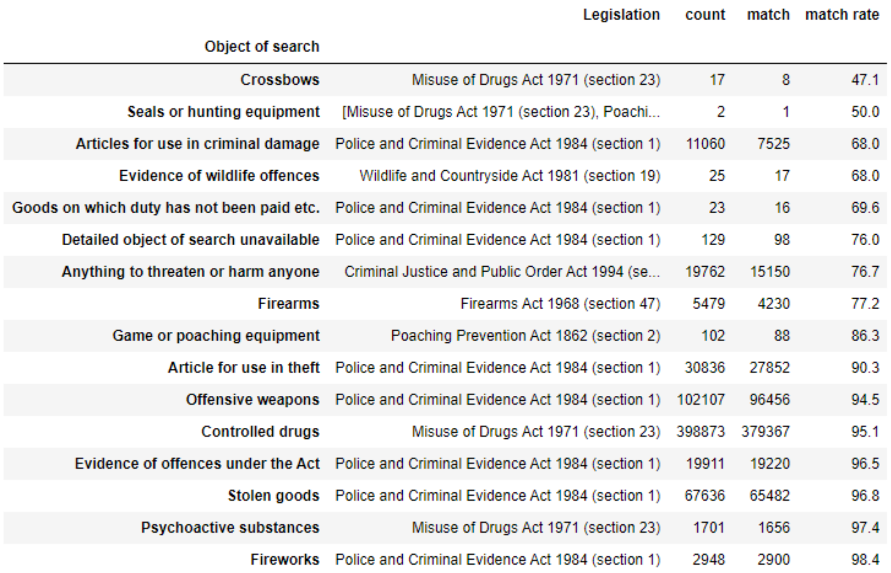

count
0
50000
100000
150000
200000
223397
10665
False
True
Removal of more than just outer clothing


count
400000
300000
200000
100000
0
455342
94118
41311
17173
16188
12232
11214
4649
3179
2102
1333
636
524
414
189
7
A no further action disposal
Arrest
Community resolution
Khat or Cannabis warning
Nothing found - no further action
Summons / charged by post
Penalty Notice for Disorder
Suspect arrested
Caution (simple or conditional)
Offender given drugs possession warning
Local resolution
Suspect summonsed to court
Article found - Detailed outcome unavailable
Offender given penalty notice
Offender cautioned
Suspected psychoactive substances seized - No further action
Outcome


count
100000
80000
60000
40000
20000
0
100273
87238
True
False
Outcome linked to object of search

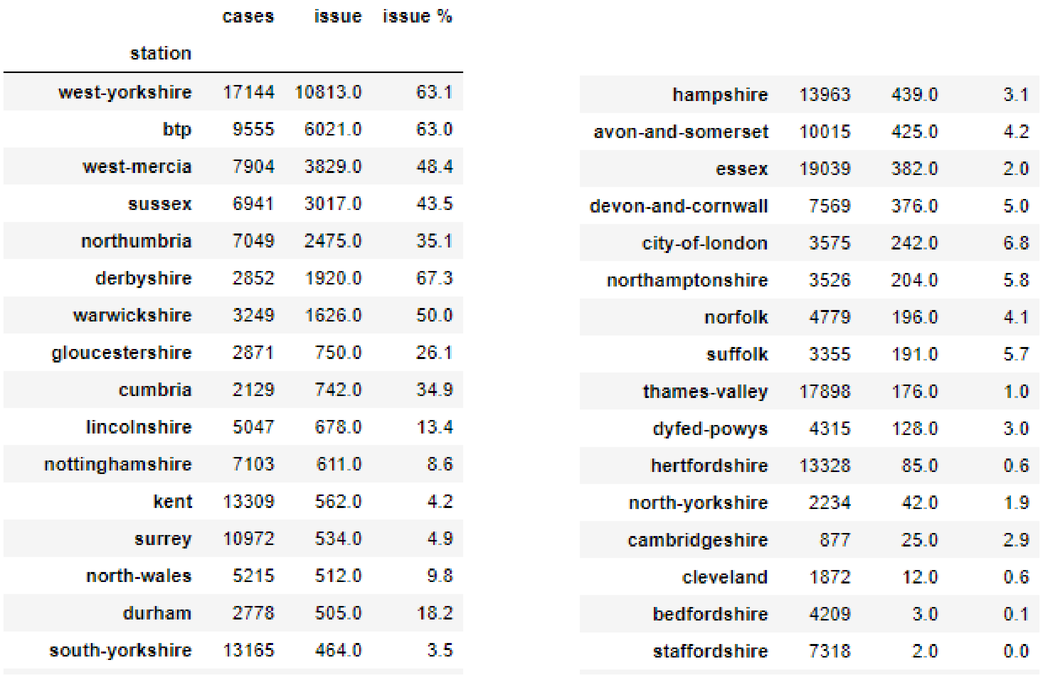
Business questions technical support

We used the correlation between Legislation and Object of search to fill some missing values. We can see that in 398873 search cases of ‘Controlled drugs’, in 95.1% of those cases the Legislation was the same, ‘Misuse of Drugs Act 1971 (section 23)’.



	Legislation	count	match	match rate
Object of search				
Crossbows	Misuse of Drugs Act 1971 (section 23)	17	8	47.1
Seals or hunting equipment	[Misuse of Drugs Act 1971 (section 23), Poachi...	2	1	50.0
Articles for use in criminal damage	Police and Criminal Evidence Act 1984 (section 1)	11060	7525	68.0
Evidence of wildlife offences	Wildlife and Countryside Act 1981 (section 19)	25	17	68.0
Goods on which duty has not been paid etc.	Police and Criminal Evidence Act 1984 (section 1)	23	16	69.6
Detailed object of search unavailable	Police and Criminal Evidence Act 1984 (section 1)	129	98	76.0
Anything to threaten or harm anyone	Criminal Justice and Public Order Act 1994 (se...	19762	15150	76.7
Firearms	Firearms Act 1968 (section 47)	5479	4230	77.2
Game or poaching equipment	Poaching Prevention Act 1862 (section 2)	102	88	86.3
Article for use in theft	Police and Criminal Evidence Act 1984 (section 1)	30836	27852	90.3
Offensive weapons	Police and Criminal Evidence Act 1984 (section 1)	102107	96456	94.5
Controlled drugs	Misuse of Drugs Act 1971 (section 23)	398873	379367	95.1
Evidence of offences under the Act	Police and Criminal Evidence Act 1984 (section 1)	19911	19220	96.5
Stolen goods	Police and Criminal Evidence Act 1984 (section 1)	67636	65482	96.8
Psychoactive substances	Misuse of Drugs Act 1971 (section 23)	1701	1656	97.4
Fireworks	Police and Criminal Evidence Act 1984 (section 1)	2948	2900	98.4

We noticed a possible issue, there are records with a ‘True’ ‘Outcome linked to object of search’ and an ‘Outcome’ of ‘Nothing found - no further action’ or ‘A no further action disposal’.



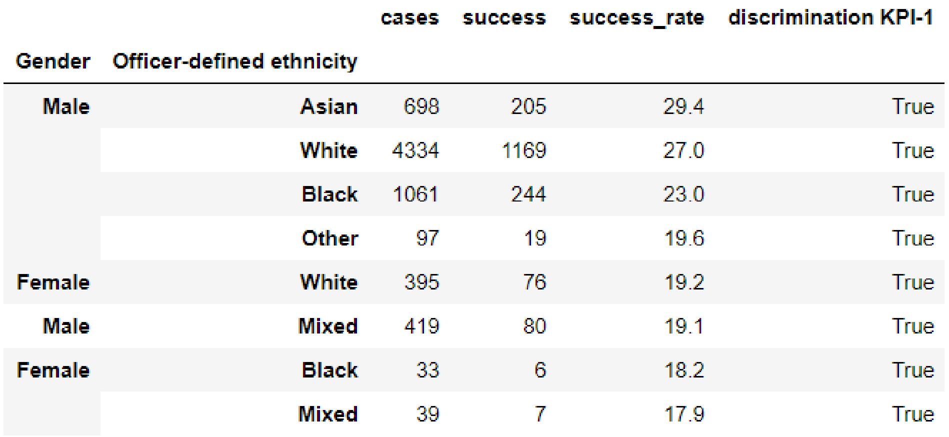
	cases	issue	issue %
station			
west-yorkshire	17144	10813.0	63.1
btp	9555	6021.0	63.0
west-mercia	7904	3829.0	48.4
sussex	6941	3017.0	43.5
northumbria	7049	2475.0	35.1
derbyshire	2852	1920.0	67.3
warwickshire	3249	1626.0	50.0
gloucestershire	2871	750.0	26.1
cumbria	2129	742.0	34.9
lincolnshire	5047	678.0	13.4
nottinghamshire	7103	611.0	8.6
kent	13309	562.0	4.2
surrey	10972	534.0	4.9
north-wales	5215	512.0	9.8
durham	2778	505.0	18.2
south-yorkshire	13165	464.0	3.5
hampshire	13963	439.0	3.1
avon-and-somerset	10015	425.0	4.2
essex	19039	382.0	2.0
devon-and-cornwall	7569	376.0	5.0
city-of-london	3575	242.0	6.8
northamptonshire	3526	204.0	5.8
norfolk	4779	196.0	4.1
suffolk	3355	191.0	5.7
thames-valley	17898	176.0	1.0
dyfed-powys	4315	128.0	3.0
hertfordshire	13328	85.0	0.6
north-yorkshire	2234	42.0	1.9
cambridgeshire	877	25.0	2.9
cleveland	1872	12.0	0.6
bedfordshire	4209	3.0	0.1
staffordshire	7318	2.0	0.0

To better understand the claims against the Department, we created 6 KPIs to evaluate available data.

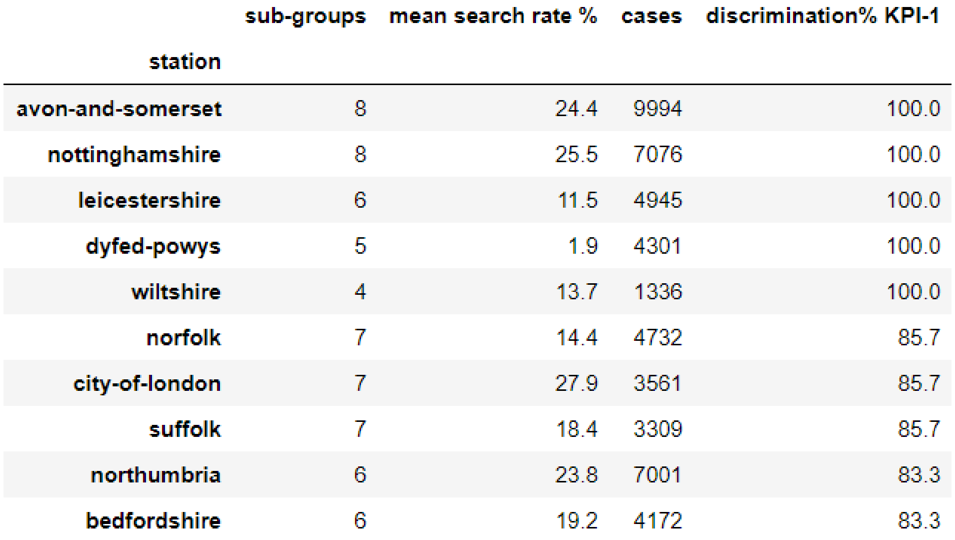
KPI-1 (station indicator)

- Evaluates discrimination using a success rate criteria per sub-group in each station
- ‘True’ means that the sub-group fails the discrimination criteria (is being discriminated)
- ‘100%’ means that all sub-groups fail the criteria

Below we present the top10 stations with the highest index of discrimination and an example station. In ‘nottinghamshire’ we have a mean success search rate of 25.5% with 8 sub-groups and all sub-groups follow outside the +/-5% margin thus, all sub-groups are considered as a discriminated sub-group (True). Then, we considered for this station a KPI-1 of 100% (8 out of 8 sub-groups).



		cases	success	success_rate	discrimination KPI-1
Gender	**Officer-defined ethnicity**				
Male	**Asian**	698	205	29.4	True
	White	4334	1169	27.0	True
	Black	1061	244	23.0	True
	Other	97	19	19.6	True
Female	**White**	395	76	19.2	True
Male	**Mixed**	419	80	19.1	True
Female	**Black**	33	6	18.2	True
	Mixed	39	7	17.9	True

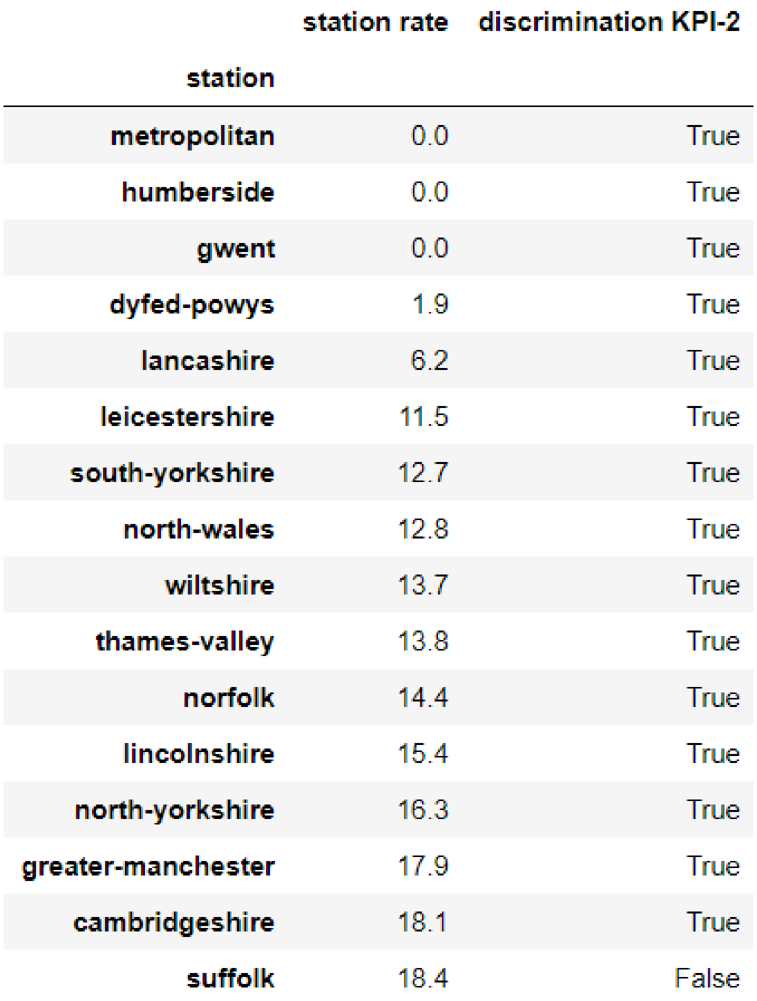


	sub-groups	mean search rate %	cases	discrimination% KPI-1
station				
avon-and-somerset	8	24.4	9994	100.0
nottinghamshire	8	25.5	7076	100.0
leicestershire	6	11.5	4945	100.0
dyfed-powys	5	1.9	4301	100.0
wiltshire	4	13.7	1336	100.0
norfolk	7	14.4	4732	85.7
city-of-london	7	27.9	3561	85.7
suffolk	7	18.4	3309	85.7
northumbria	6	23.8	7001	83.3
bedfordshire	6	19.2	4172	83.3

KPI-2 (global indicator)

- Evaluates discrimination using a success rate criteria per station
- ‘True’ means that the station fails the discrimination criteria (is being discriminated)

The general overview over the Department shows that 31% of 39 stations (3 stations without success cases were removed) are below the overall mean margin and 41% are above. Bottom place in KPI-2 ranking is Dyfed-powys with a station rate of only 1.9% and Top place is Durham with a station rate of 35.8%.

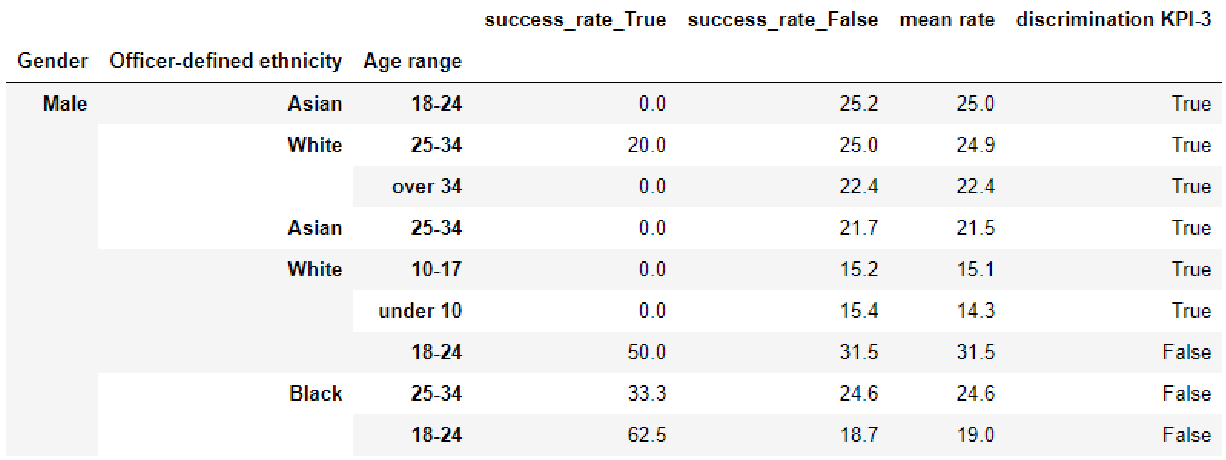


	station rate	discrimination KPI-2
station		
metropolitan	0.0	True
humberside	0.0	True
gwent	0.0	True
dyfed-powys	1.9	True
lancashire	6.2	True
leicestershire	11.5	True
south-yorkshire	12.7	True
north-wales	12.8	True
wiltshire	13.7	True
thames-valley	13.8	True
norfolk	14.4	True
lincolnshire	15.4	True
north-yorkshire	16.3	True
greater-manchester	17.9	True
cambridgeshire	18.1	True
suffolk	18.4	False

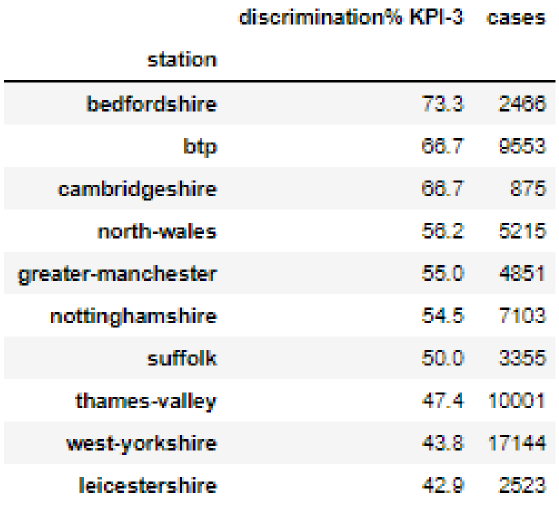
KPI-3 (station indicator)

- Evaluates discrimination to remove clothes using a success rate criteria per sub-group in each station
- ‘True’ means that the sub-group fails the discrimination criteria (is being discriminated)
- ‘73.3%’ is the percentage of sub-groups available for that station which fail the criteria

In btp (station) we have mean success search rates (per sub-group) from 14.3% to 31.5% and we evaluated each sub-group if the success rate when Officer asked to remove clothes is not lower than the mean rate with a 10% margin. There are 6 sub-groups with a significant drop in search rate so we considered that station btp has a discrimination KPI-3 of 66.7% (6 out of 9) and these 6 sub-groups will be counted in KPI-4. We also present the top10 station ranking of KPI-3.



			success_rate_True	success_rate_False	mean rate	discrimination KPI-3
Gender	Officer-defined ethnicity	Age range				
Male	Asian	18-24	0.0	25.2	25.0	True
	White	25-34	20.0	25.0	24.9	True
		over 34	0.0	22.4	22.4	True
	Asian	25-34	0.0	21.7	21.5	True
	White	10-17	0.0	15.2	15.1	True
		under 10	0.0	15.4	14.3	True
		18-24	50.0	31.5	31.5	False
	Black	25-34	33.3	24.6	24.6	False
		18-24	62.5	18.7	19.0	False

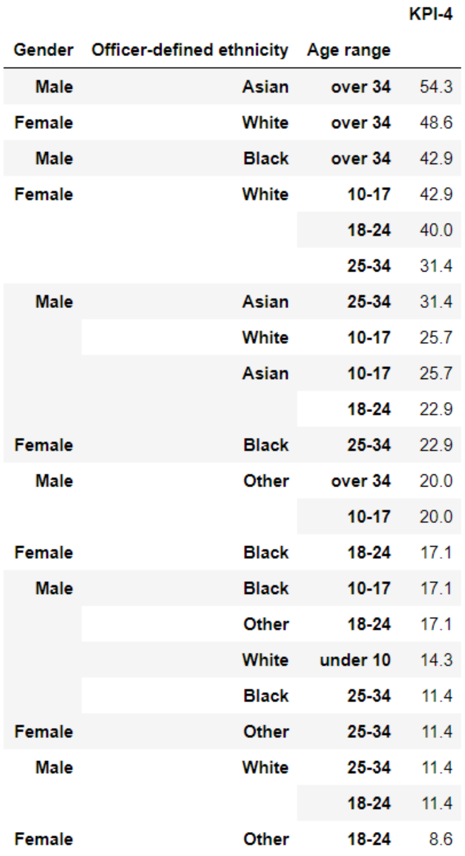


	discrimination% KPI-3	cases
station		
bedfordshire	73.3	2468
btp	66.7	9553
cambridgeshire	66.7	875
north-wales	56.2	5215
greater-manchester	55.0	4851
nottinghamshire	54.5	7103
suffolk	50.0	3355
thames-valley	47.4	10001
west-yorkshire	43.8	17144
leicestershire	42.9	2523

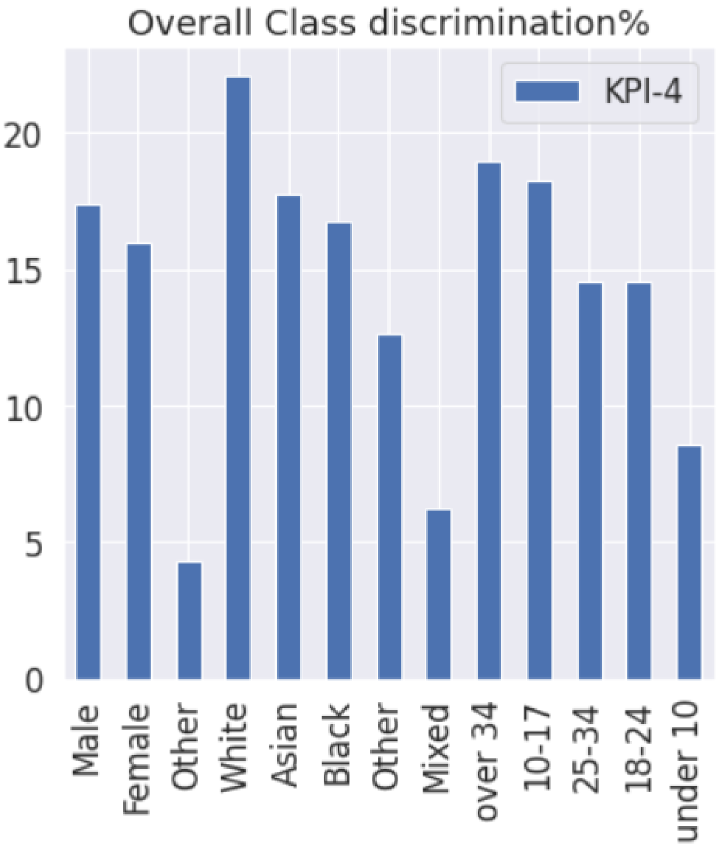
KPI-4 (global indicator)

- Aggregates the results from KPI-3 sub-groups for all stations
- ‘54.3%’ is the percentage of stations that are discriminating that sub-group
- Overall Class discrimination represents the percentage discrimination per category

Globally, KPI-4 ranking reveals which sub-groups are being discriminated to remove clothes in more stations. An asian male over 34 years is discriminated in 54.3% of all 35 stations with valid data. Looking at classes separately, ‘Male’, ‘White’ and ‘over 34’ are more predominant in more stations.



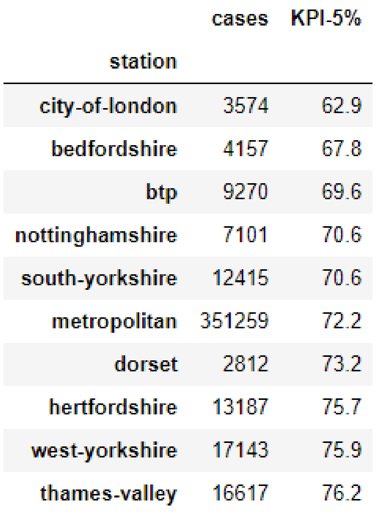
			KPI-4
Gender	**Officer-defined ethnicity**	**Age range**	
Male	**Asian**	**over 34**	54.3
Female	**White**	**over 34**	48.6
Male	**Black**	**over 34**	42.9
Female	**White**	**10-17**	42.9
		18-24	40.0
		25-34	31.4
Male	**Asian**	**25-34**	31.4
	White	**10-17**	25.7
	Asian	**10-17**	25.7
		18-24	22.9
Female	**Black**	**25-34**	22.9
Male	**Other**	**over 34**	20.0
		10-17	20.0
Female	**Black**	**18-24**	17.1
Male	**Black**	**10-17**	17.1
	Other	**18-24**	17.1
	White	**under 10**	14.3
	Black	**25-34**	11.4
Female	**Other**	**25-34**	11.4
Male	**White**	**25-34**	11.4
		18-24	11.4
Female	**Other**	**18-24**	8.6



KPI-5 (station indicator)

- Evaluates discrimination (mismatch) between ‘Self-defined ethnicity’ and ‘Officer-defined ethnicity’ in each station
- ‘62.9%’ is the percentage of cases in which the definition of the agent matches the suspect’s

Our ranking indicates that Officers from City-of-london, Bedfordshire and Btp do not accurately match suspect’s self-defined ethnicity 70% of occurrences, in other words, they might be discriminating 30% of the suspects.

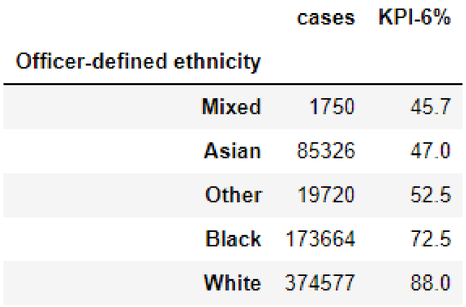


	cases	KPI-5%
station		
city-of-london	3574	62.9
bedfordshire	4157	67.8
btp	9270	69.6
nottinghamshire	7101	70.6
south-yorkshire	12415	70.6
metropolitan	351259	72.2
dorset	2812	73.2
hertfordshire	13187	75.7
west-yorkshire	17143	75.9
thames-valley	16617	76.2

KPI-6 (global indicator)

- Evaluates discrimination using the match percentage of cases between ‘Self-defined ethnicity’ and ‘Officer-defined ethnicity’
- ‘45.7%’ is the percentage of cases that Officer’s definition match suspect’s

On a global perspective, the Department has a reasonable accuracy for White and Black ethnies but considerably lower accuracy for all other ethnies.



	cases	KPI-6%
Officer-defined ethnicity		
Mixed	1750	45.7
Asian	85326	47.0
Other	19720	52.5
Black	173664	72.5
White	374577	88.0

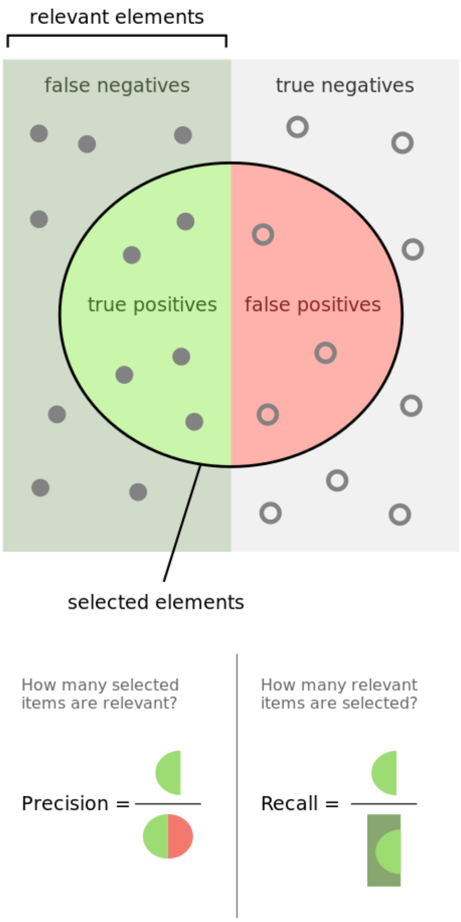
Model technical analysis

Our model is divided into 3 groups of features and 1 classifier :

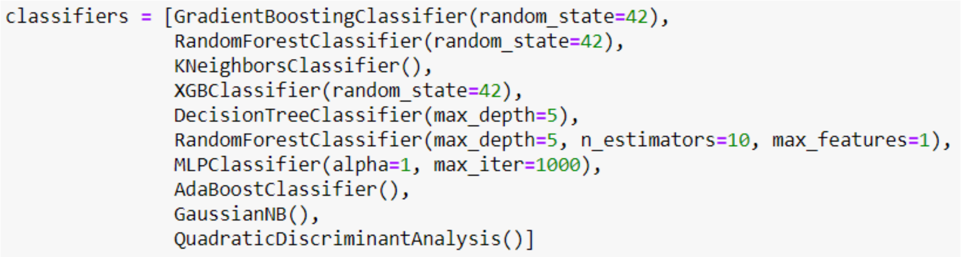
1. ‘Part of a policing operation’, ‘Hour’, ‘Week_day’, ‘Object of search’, ‘Officer-defined’, ‘ethnicity', ‘Age range’, ‘Gender’ and ‘Type’
    a. SimpleImputer with ‘strategy’ equals to ‘most_frequent’
    b. OneHotEncoder with ‘handle_unknown’ equals to ‘ignore’
2. ‘station’
    a. SimpleImputer with ‘strategy’/’fill_value’ equals to ‘constant’/’None’
    b. TargetEncoder
3. ‘Latitude’ and ‘Longitude’
    a. SimpleImputer with ‘strategy’ equals to ‘mean’
    b. KBinsDiscretizer with ‘n_bins’/’encode’/’strategy’ equals to ‘10’/’onehot’/’kmeans’
4. Classifier
    a. GradientBoostingClassifier with ‘random_state’ equals to 42

All other parameters of our model are the default ones from scikit-learn.

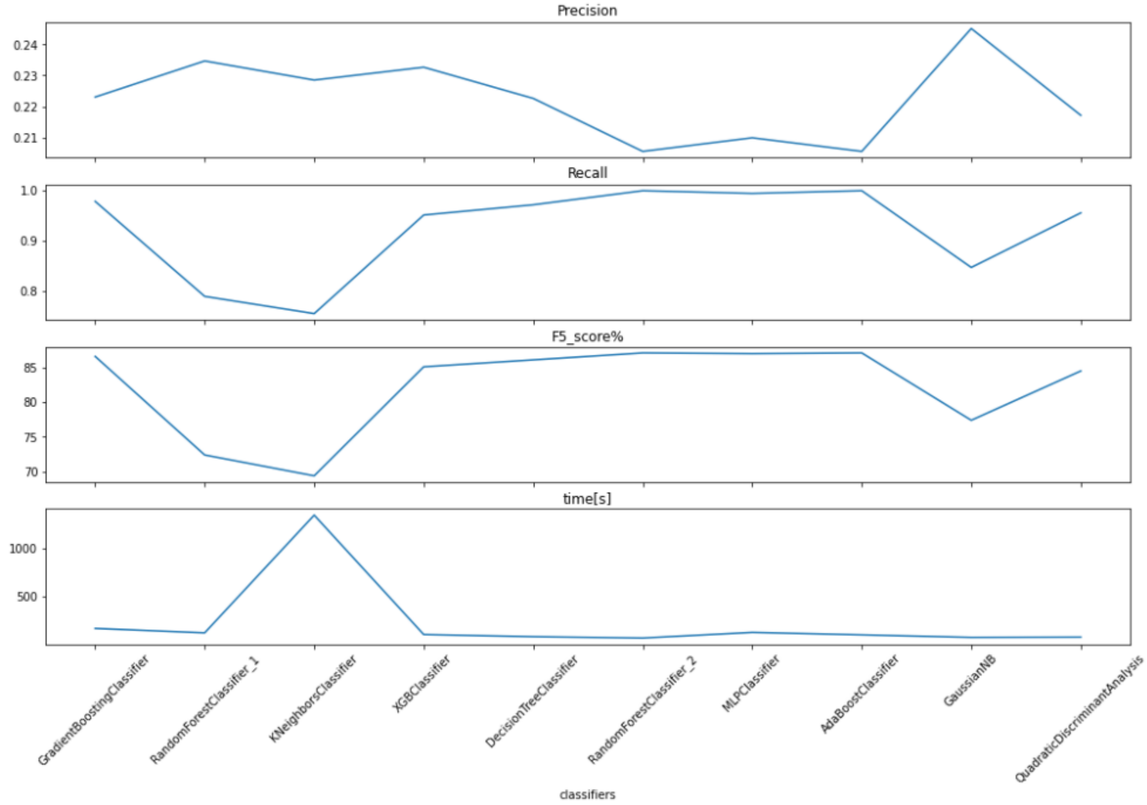
To better understand some results and KPIs presented in this report, please consider the following illustration.

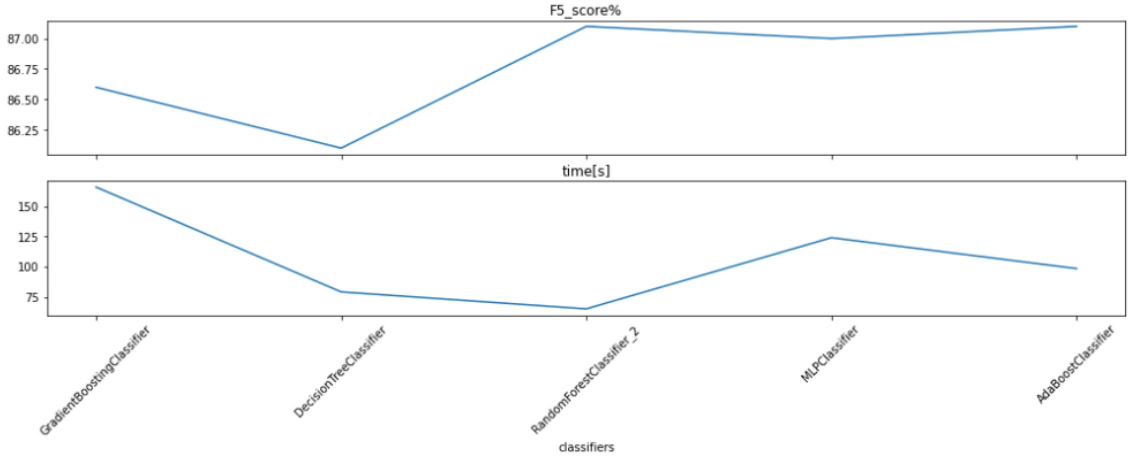


Alternative classifiers were tested with the following parameters (mostly default parameters):

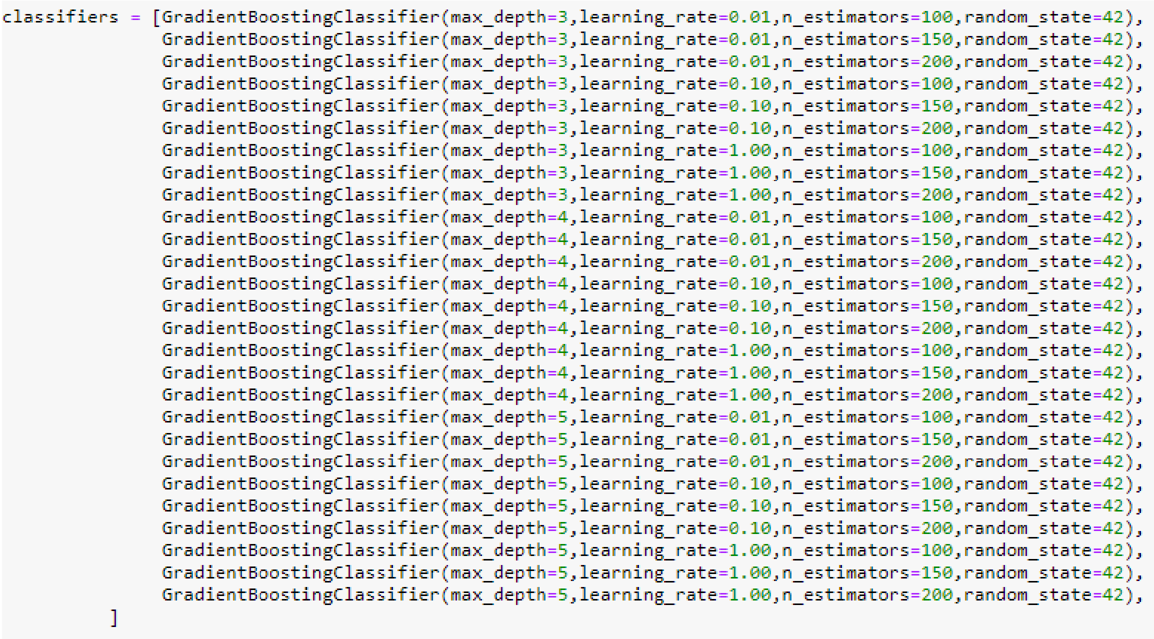


```
classifiers = [GradientBoostingClassifier(random_state=42),
               RandomForestClassifier(random_state=42),
               KNeighborsClassifier(),
               XGBClassifier(random_state=42),
               DecisionTreeClassifier(max_depth=5),
               RandomForestClassifier(max_depth=5, n_estimators=10, max_features=1),
               MLPClassifier(alpha=1, max_iter=1000),
               AdaBoostClassifier(),
               GaussianNB(),
               QuadraticDiscriminantAnalysis()]
```

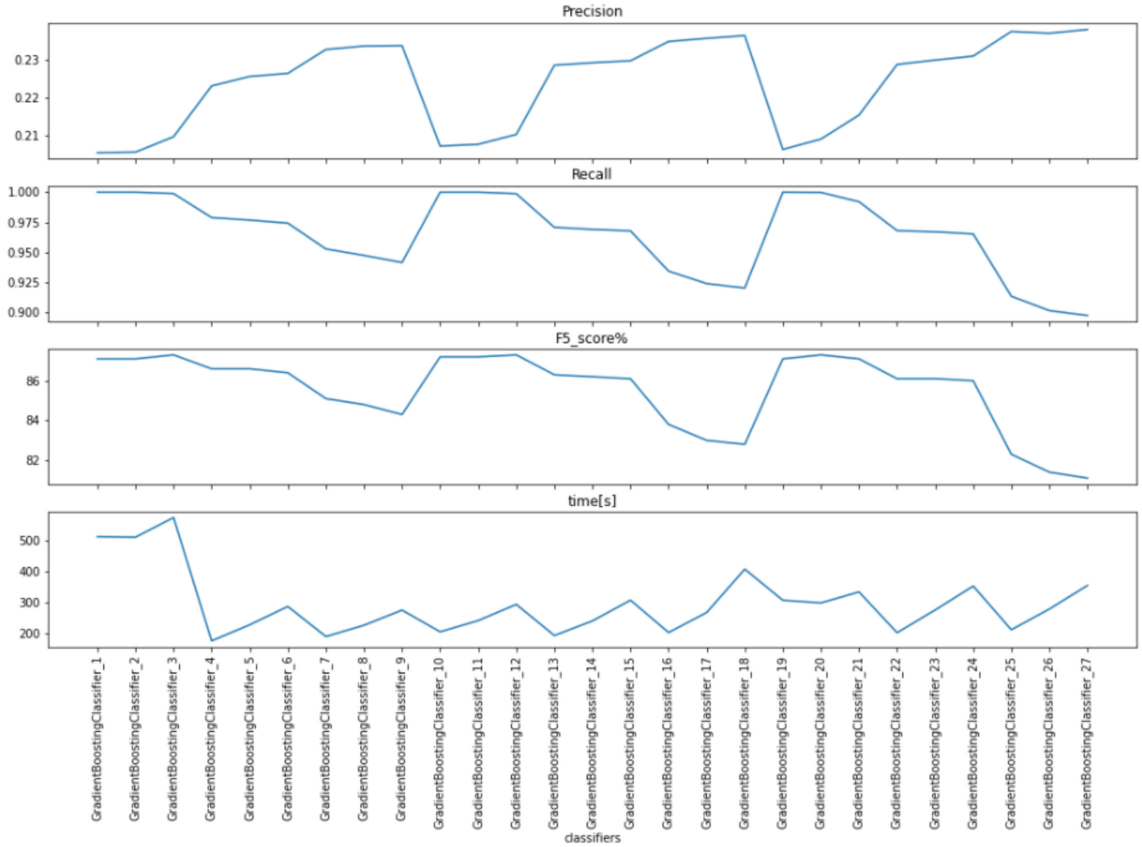




Parameter tuning results for 'GradientBoostingClassifier':



```
classifiers = [GradientBoostingClassifier(max_depth=3,learning_rate=0.01,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=3,learning_rate=0.01,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=3,learning_rate=0.01,n_estimators=200,random_state=42),
               GradientBoostingClassifier(max_depth=3,learning_rate=0.10,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=3,learning_rate=0.10,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=3,learning_rate=0.10,n_estimators=200,random_state=42),
               GradientBoostingClassifier(max_depth=3,learning_rate=1.00,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=3,learning_rate=1.00,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=3,learning_rate=1.00,n_estimators=200,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=0.01,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=0.01,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=0.01,n_estimators=200,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=0.10,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=0.10,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=0.10,n_estimators=200,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=1.00,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=1.00,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=4,learning_rate=1.00,n_estimators=200,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=0.01,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=0.01,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=0.01,n_estimators=200,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=0.10,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=0.10,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=0.10,n_estimators=200,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=1.00,n_estimators=100,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=1.00,n_estimators=150,random_state=42),
               GradientBoostingClassifier(max_depth=5,learning_rate=1.00,n_estimators=200,random_state=42),
          ]
```


Precision
Recall
F5_score%
time[s]
GradientBoostingClassifier_1
GradientBoostingClassifier_2
GradientBoostingClassifier_3
GradientBoostingClassifier_4
GradientBoostingClassifier_5
GradientBoostingClassifier_6
GradientBoostingClassifier_7
GradientBoostingClassifier_8
GradientBoostingClassifier_9
GradientBoostingClassifier_10
GradientBoostingClassifier_11
GradientBoostingClassifier_12
GradientBoostingClassifier_13
GradientBoostingClassifier_14
GradientBoostingClassifier_15
GradientBoostingClassifier_16
GradientBoostingClassifier_17
GradientBoostingClassifier_18
GradientBoostingClassifier_19
GradientBoostingClassifier_20
GradientBoostingClassifier_21
GradientBoostingClassifier_22
GradientBoostingClassifier_23
GradientBoostingClassifier_24
GradientBoostingClassifier_25
GradientBoostingClassifier_26
GradientBoostingClassifier_27
classifiers